Structure Preserving Spectral Methods and Exponential Integrators for the Numerical Solution of Stiff Semi-Linear Partial Differential Equations

by

Emmanuel Appiah-Kubi

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Mathematics
University of Manitoba
Winnipeg

Abstract

In this thesis, we present numerical solution of semilinear partial differential equations (PDEs) where the linear differential operator is a self-adjoint. A recent spectral method for self-adjoint operators, based on basis recombination, leads to symmetric definite matrices, which have real spectrum. This allows for developing stable time-stepping algorithm to solve the resulting the ordinary differential equations.

The linear part of the discretized problem is usually stiff, which constraints the step size for explicit numerical schemes. We therefore use exponential integrators, a well-known time-stepping methods for solving stiff differential equations. We describe three methods namely, eigen-decomposition, contour integral and Carathéodory-Fejér approximation, for computing the matrix (φ) functions of the exponential integrators.

We perform numerical experiments with some PDEs with different boundary conditions, including time-dependent boundary condition and the numerical results confirm the accuracy of the methods in both space and time.

Acknowledgements

First of all, I thank the Almighty God for His provisions, good health and strength throughout this journey. Indeed, "...victory cometh from the LORD"

A profound gratitude goes to my supervisor Dr Richard Mikael Slevinsky for his immense contribution towards this work. His constructive criticism, motivation, patience and guidance helped to make this work a success. He was always available for discussions. In fact, it was a great pleasure to work under his supervision.

The next appreciation goes to the Department of Mathematics for the funding and also the resourceful faculty who contributed to my studies in diverse ways. I also appreciate my thesis committee Dr Shaun Lui and Dr Craig Cowan for their valuable comments and remarks that helped to improve the thesis. I'm grateful to Dr Leo Butler for his advice and guidance.

A very big thanks to my family especially my mother Georgina Fosu Agyeiwaa, my father Ohene Appiah-Kubi and my brother Francis for their love and support.

I am also grateful to the Ghana community in Winnipeg particularly Anderson, Clifford and Gifty for the social support and encouragement.

Contents

List of Figures

1

Introduction

1.1 Introduction

Many problems in science and engineering are of the form of time dependent semilinear partial differential equations (PDEs) of the form

$$u_t = \mathcal{L}u + \mathcal{N}(u,t), \tag{1.1}$$

where $\mathcal{L}$ is a linear differential operator and $\mathcal{N}$ a nonlinear operator.

Examples of such equations include those attributed to Allen–Cahn, Cahn–Hilliard, Gierer–Meinhardt, Ginzburg–Landau, Kuramoto–Sivashinsky, etc.... We consider the one-dimensional spatial form of these equations.

The spatial discretization of (1.1) leads to a system of ordinary differential equations (ODEs) which is mostly characterized by *stiffness* in the linear part. The stiffness leads to constraint on the step size where explicit time-stepping schemes require an extremely small time step to ensure the truncated spectrum is contained in the linear stability domain. Implicit methods can be used to alleviate this difficulty but computing the solution at the next time step can be expensive due to the nonlinearity.

There are a number of existing schemes which can handle the issue of stiffness and the well-known of this class of methods include implicit-explicit (IMEX), integrating factor scheme, split step and exponential integrators, which is the main tool used in this thesis. Exponential integrators have been developed for solving semilinear PDEs by solving the linear part exactly and then "approximating" the nonlinear part of the system. This results in computing the matrix exponentials and matrix functions (the so-called φ functions), which will be discussed in Chapter 2. Numerical experiments conducted by Kassam and Trefethen [2005], Montanelli and Bootland [2016] confirm the superiority of exponential integrators over other existing methods for solving stiff semilinear PDEs.

In this thesis, we are interested in solving (1.1) when the linear operator $\mathcal{L}$ is a formally self-adjoint linear differential eigenvalue equipped self-adjoint boundary conditions defined as

$$\mathcal{L}u = \lambda u, \qquad \mathcal{B}u = 0, \tag{1.2}$$

where $\mathcal{L}$ is a self-adjoint linear differential operator of the form

$$\mathcal{L} = (-\mathcal{D})^N \left(p_N \mathcal{D}^N\right) + (-\mathcal{D})^{N-1} \left(p_{N-1} \mathcal{D}^{N-1}\right) + \cdots + p_0, \tag{1.3}$$

where the variable coefficients are polynomials of degree at most m, $p_i \in \mathbb{P}_m$, and $p_N \neq 0$.

The spatial discretization using the well-known Ritz–Galerkin method and subsequently representing the solution as a linear combination of orthogonal polynomials leads to a symmetric but dense matrix and hence higher computational cost (possibly $\mathcal{O}(N^3)$). A modern spectral method is "the ultraspherical spectral method" developed by Olver and Townsend [2013] seeks to represent the solution of a linear differential equations (with variable coefficients) in a Chebyshev polynomial expansion by making use of the relationship between derivatives of Chebyshev polynomials and ultraspherical polynomials. The boundary conditions are imposed using respective number of rows of the linear system. The resulting discretization matrix is banded but non-symmetric even for self-adjoint differential equations.

More recently, Aurentz and Slevinsky [2020] showed that the symmetry and sparsity in self-adjoint differential equations with polynomial coefficients may transcend the ultraspherical spectral discretization. This uses a recombination of orthogonal polynomial bases and leads to symmetric-definite and banded discretization matrices. Symmetry guarantees real eigenvalues and eigenvector can be chosen to be orthogonal to each other.

The symmetric-definite discretizations are important because they permit stable numerical algorithms to be implemented for the generalized eigendecomposition. The banded structure results in a significant improvement in the complexity of the stable algorithms which is reduced from from the general dense case of $\mathcal{O}(N^3)$ down to $\mathcal{O}(mN^2)$, where the bandwidth is $\mathcal{O}(m)$. In this work, we use fixed-size spatial discretizations. It is known Aurentz and Slevinsky [2020] that principal finite sections of ultraspherical spectral discretizations do not respect the symmetry of self-adjoint linear differential operators. Therefore, the fact that the true spectrum and that of any of its principal finite sections of the method of Aurentz and Slevinsky [2020] is real is crucial in that it permits the design and implementation of contour integral representations of functions of the linear operator and rational functions more generally.

1.1.1 Second Order Linear Eigenvalue Differential Problem

Considering the classical negative second order linear eigenvalue differential equation with boundary condition which is a well-known self-adjoint problem

$$-\mathcal{D}^2 u = \lambda u, \quad \mathcal{B}u = 0. \tag{1.4}$$

For a separable Hilbert space $H = L^2([-1,1])$ of the eigenfunctions, we define a quotient space which satisfies the boundary condition as $H_{\mathcal{B}} = \{u \in H : \mathcal{B}u = 0\}$. We represent the eigenfunction in normalized weighted Jacobi polynomial expansion

$$u(x) = \sum_{n=0}^{\infty} u_n (1 - x^2) \tilde{P}_n^{(1,1)}(x). \tag{1.5}$$

With the appropriate choice of basis in the quotient space and following the Aurentz–Slevinsky method which will be discussed in Chapter 2, we end up with the discretized form of (1.4) as

$$A\mathbf{u} = \lambda B\mathbf{u}, \tag{1.6}$$

where A is diagonal and B is symmetric pentadiagonal and positive definite pentadiagonal matrix. Here λ is the generalized eigenvalue of the pencil (A, B) and $\mathbf{u}$ is the coefficient of the expansion of the solution of the problem (1.4).

We relate this problem to the PDE defined in (1.1) where the discretized linear part becomes $L = -B^{-1}A$ for the negative second order linear differential operator in Eq (1.4).

We consider a classical one-dimensional equation of the form (1.1)

$$u_t = \epsilon u_{xx} + u - u^3, \quad x \in [-1,1] \quad t \geq 0 \tag{1.7}$$

known as the Allen–Cahn equation (Cahn and Allen [1977]), which is a well-known reaction diffusion equation used to model phase-separation of alloys in material science. It has a second order diffusive term with a cubic reaction term and we consider the equation in aperiodic domain $x \in [-1, 1]$.

Kassam and Trefethen [2005] solved this equation using Cox and Matthews [2002] fourth-order exponential integrators by first discretizing the PDE with a Chebyshev pseudo-spectral method, leading to a dense non-symmetric matrix for the linear part of (1.7). Due to the fact that they worked with a dense matrix, the computation of the matrix (φ) functions exponential integrators using contour integral on a unit circle and subsequent precomputations results in $\mathcal{O}(N^4)$ complexity with time-stepping complexity of $\mathcal{O}(N^3)$ from matrix-vector product.

The objective of this project is to improve the computational complexity particularly for self-adjoint problems using the following approaches to compute the matrix functions

1. eigen-decomposition of the linear part resulting from the Aurentz and Slevinsky [2020] method,
2. the use of contour integral for the computation of the φ functions (Hale and Weideman [2015], Weideman and Trefethen [2007]),
3. rational approximation of the φ functions, specifically the Carathéodory–Fejér method (Trefethen and Schmelzer [2007])

Different boundary conditions will also be considered including nonhomogeneous and time-dependent boundary conditions and we will also look at implementing it on other PDEs of the form (1.1).

We consider self-adjoint linear differential operators in this research due to their numerous applications in applied mathematics, particularly in mathematical physics. The nonlinearity in 1.1 depends on the particular semilinear PDE.

2

Background Research

2.1 Self-Adjoint Linear Differential Operators

Definition 2.1. Kufner [1985], Lui [2012] Let $\Omega \subset \mathbb{R}^d$ and define $w : \mathbb{R}^d \to [0, \infty)$ as the weight function. We define the weighed Sobolev space for every $1 \leq p < \infty$ and for every $m \in \mathbb{N}_0$ as

$$W^{m,p}(\Omega, w) = \{u \in L^p(\Omega, w), \partial^\alpha u \in L^p(\Omega, w), \forall \alpha \in \mathbb{N}^d, |\alpha| \leq m\},$$

endowed with the norm

$$||u||_{W^{m,p}(\Omega,w)} = \left(\sum_{|\alpha| \leq m} \int_\Omega |\partial^\alpha u|^p w(x) dx \right)^{1/p}.$$

For $p = 2$, we denote $W^{m,2}(\Omega, w) = H^m_w(\Omega)$.

Let $H = L^2([a, b], w(x)dx)$ be Hilbert space of square-integrable functions on $[a, b]$. We define an inner product as

$$\langle f, g \rangle = \int_a^b \overline{f(x)} g(x) w(x) dx,$$

for $f, g \in H$ and $\overline{f(x)}$ denotes the complex conjugate of $f(x)$. The adjoint of an operator $\mathcal{L}$ denoted $\mathcal{L}^*$ is defined (Boyce and DiPrima [1977], Hall [2013], Stone and Goldbart [2009]) such that

$$\langle f, \mathcal{L}g \rangle = \langle \mathcal{L}^* f, g \rangle.$$

If $\mathcal{L} = \mathcal{L}^*$, then $\mathcal{L}$ is said to be *self-adjoint*.

A linear differential operator that is self-adjoint when equipped with appropriate boundary conditions is called formally self-adjoint when those boundary conditions are absent.

Some examples of formally self-adjoint linear differential operators appearing in eigenvalue problems include:

- The Sturm–Liouville eigenvalue problem

  $$(-\mathcal{D})\,(p\mathcal{D})\,u + qu = \lambda w u, \tag{2.1}$$

  where $p \in C^1([a, b])$, $q \in C([a, b])$, and $0 < w \in C([a, b])$. The second order operator has many of applications in mathematical physics, specifically in electrostatics, vibrating string, and quantum theory Stone and Goldbart [2009].

- Choosing $p(x) = 1$, $q(x) = 0$ and $w(x) = 1$ in Eq. (2.1) leads to the (negative) Laplace eigenvalue problem

  $$-\mathcal{D}^2 u = \lambda u.$$

- The momentum operator given by

  $$i\mathcal{D}u = \lambda u,$$

is also another formally self-adjoint operator used in quantum theory (Hall [2013]).

We prove two results of self-adjoint operators which will be useful in this thesis.

Theorem 2.2. (Griffel [2002]) Eigenvalues of self-adjoint operators are real.

Proof. Let u be an eigenfunction corresponding to the eigenvalue λ, that is $\mathcal{L}u = \lambda u$, then we have

$$\begin{aligned} \lambda\langle u, u\rangle &= \langle u, \lambda u\rangle, \\ &= \langle u, \mathcal{L}u\rangle, \\ &= \overline{\langle \mathcal{L}u, u\rangle}, \\ &= \overline{\lambda}\langle u, u\rangle. \end{aligned}$$

Since $\langle u, u\rangle \neq 0$, we have $\lambda = \overline{\lambda}$. Therefore λ is real. □

Theorem 2.3. (Griffel [2002]) If λ and μ are distinct eigenvalues of a self-adjoint operator with eigenfunctions u and v respectively, then u and v are orthogonal.

Proof. From the self-adjoint property, we have

$$\begin{aligned} \langle \mathcal{L}u, v\rangle &= \langle u, \mathcal{L}v\rangle \\ \implies \lambda\langle u, v\rangle &= \mu\langle u, v\rangle \\ (\lambda - \mu)\langle u, v\rangle &= 0 \end{aligned}$$

Since $\lambda \neq \mu$, we have $\langle u, v\rangle = 0$. Thus the eigenfunction corresponding to different eigenvalues are orthogonal. □

2.2 Exponential Integrators

Consider the initial-value problem

$$y'(t) = f(t, y), \qquad y(t_0) = y_0. \tag{2.2}$$

Given a step size $h > 0$, a time-stepping scheme produces a sequence $\{y_n\}_{n=0}^{\infty}$ that approximates the true solution $y(t_n)$ at time $t_n = t_0 + nh$.

2.2.1 Linear Stability Analysis

Consider the linear scalar initial-value problem

$$y'(t) = \lambda y(t), \qquad y(t_0) = y_0, \quad \lambda \in \mathbb{C}, \tag{2.3}$$

with exact solution $y(t) = e^{\lambda t} y_0$.

Suppose that $\mathrm{Re}(\lambda) < 0$ such that the exact solution decays to zero, that is, $\lim\limits_{t \to \infty} y(t) = 0$. Thus for a fixed time step h, we define the linear stability domain as $\mathcal{D}_s$ of a numerical method as the set of all number $z = \lambda h \in \mathbb{C}$ such that $\lim_{n \to \infty} y_n = 0$ (Iserles [2009]). A method is termed as Absolutely-stable (A-stable) if the linear stability domain contains the set $h\lambda \in \mathbb{C} : \mathrm{Re}(h\lambda) < 0$ (Dahlquist [1963]).

2.2.2 Derivation of Exponential integrators

Exponential integrators have been developed to solve semilinear PDEs of the form (1.1) with stiffness in the linear part. Stiffness is a general term used to describe the situation where the ratio of the largest to the smallest eigenvalues of a linear vector-valued initial-value problem is large. Stiffness in the discretized form of (1.1) can be alleviated by implicit methods but they become computationally expensive. Explicit methods turn out to be constrained by small stability regions, requiring extremely small time steps.

The idea behind exponential integrators is to find exact solution of the linear part and numerical approximation for the nonlinear part. This leads to applying the matrix exponential and other related matrix functions to a vector, thus the name *exponential integrators*, according to a comprehensive review done by Hochbruck and Ostermann [2010], Cox and Matthews [2002].

A review paper by Minchev and Wright [2005] indicates that different forms of exponential integrators have been rediscovered in different ways. The family of exponential integrators consists of exponential time differencing (ETD) which is subcategorized into one-stage schemes (Runge–Kutta types) and multi-stage (Adams–Bashforth), integrating factor (IF), Lawson and exponential predictor-corrector methods.

The exponential time differencing (ETD) was derived by Cox and Matthews [2002] using the integrating factor(IF) idea.

Consider a system of ODEs resulting from the semi-discretization of Eq. (1.1) as

$$u_t = Lu + N(u,t), \tag{2.4}$$

where L is the linear operator with stiffness and $N(u,t)$ a nonlinear forcing term. For a scalar case, L is a constant. ETD schemes are derived by multiplying Eq. (2.4) by the integrating factor e^{-Lt} as

$$e^{-Lt}u_t = e^{-Lt}Lu + e^{-Lt}N(u,t), \tag{2.5}$$

and integrating over a single time step from $t = t_n$ to $t = t_{n+1} = t_n + h$ for step size h yields

$$u(t_{n+1}) = e^{Lh}u(t_n) + e^{Lh}\int_0^h e^{-L\tau}N(u(t_n+\tau), t_n+\tau)d\tau. \tag{2.6}$$

Eq. (2.6) is the exact solution of (2.4), given an initial solution $u(t_n)$ and this formula is also known as the variation-of-constant formula (Hochbruck and Ostermann [2010]). The various forms of ETD methods depend on how one approximates second term in Eq. (2.6) based on the approximation of the nonlinear term.

The simplest case approximates N by the constant $N(u_n, t_n)$ which is denoted as N_n; and also representing the numerical solution of $u(t_n)$ as u_n, we have the numerical scheme

$$u_{n+1} = e^{Lh}u_n + L^{-1}(e^{Lh} - I)N_n, \tag{2.7}$$

as the exponential time differencing of order one (ETD1). This can also be written as

$$u_{n+1} = e^{Lh}u_n + h\varphi_1(Lh)N_n, \tag{2.8}$$

where $\varphi_1(z) = (e^z - 1)/z$ is referred to as a φ function.

2.2.3 Higher Order forms of ETDs

A second order of ETD can be derived using a linear approximation of the nonlinear function instead of a constant approximation used in the case of ETD1. Higher order and more accurate schemes can be derived using higher order polynomial approximations of the nonlinearity. Cox and Matthews [2002] derived the schemes of arbitrary order using Newton's backward divided difference approximation of $N(u(t_n + \tau), t_n + \tau)$, given the information about N at the nth and previous time steps $N_n, N_{n-1}, \ldots, N_{n-s}$. The approximation of N becomes

$$N(u(t_n + \tau), t_n + \tau) \approx G_n(t_n, \tau) = \sum_{m=0}^{s-1} (-1)^m \binom{-\tau/h}{m} \nabla^m G_n(t_n),$$

where ∇ is the backward difference operator.

Substituting $N(u(t_n+\tau), t_n+\tau) \approx G_n(t_n, \tau)$ into Eq. (2.6) and simplifying further the ETD scheme of order s as

$$u_{n+1} = e^{Lh}u_n + h\sum_{k=0}^{s-1} g_m \sum_{k=0}^{m}(-1)^k \binom{m}{k} N_{n-k}, \tag{2.9}$$

where the coefficients can be expressed as recurrence relation

$$\begin{aligned} Lhg_0 + I &= e^{Lh}, \\ Lhg_{m+1} + I &= g_m + \frac{1}{2}g_{m-1} + \frac{1}{3}g_{m-2} + \cdots + \frac{1}{m+1}g_0 = \sum_{k=0}^{m} \frac{1}{m+1-k}g_k. \end{aligned}$$

In particular, ETD1 and ETD2 have coefficients

$$g_0 = (e^{Lh} - I)L^{-1}h^{-1}, \quad \text{and} \quad g_1 = (g_0 - 1)L^{-1}h^{-1} = (e^{Lh} - I - Lh)L^{-2}h^{-2},$$

as their respective φ functions.

2.2.4 Exponential Time Differencing Runge–Kutta Methods

The ETD methods are multistep schemes which require s previous evaluations of the nonlinearity. Usually the initial condition leads to one nonlinear evaluation available and thus it is more convenient to use Runge–Kutta time stepping forms of ETD schemes. In addition to that, Runge–Kutta schemes have higher accuracy and larger stability regions. The second order form of these methods can be derived by first approximating the solution u at $t_n + h$ as

$$a_n = e^{Lh}u_n + L^{-1}(e^{Lh} - I)N_n, \tag{2.10}$$

which can also be obtained from $s = 1$ in the s-step scheme (2.9).

We then approximate the nonlinear function under the integral Eq. (2.6) in the interval $t_n \leq \tau \leq t_{n+1}$. The linear polynomial interpolant is given by

$$\begin{aligned} N &= N(u_n, t_n) + \frac{\tau - t_n}{h}(N(u_n + h, t_n + h) - N(u_n, t_n)) + \mathcal{O}(h^2), \\ &= N(u_n, t_n) + \frac{\tau - t_n}{h}(N(a_n, t_n + h) - N(u_n, t_n)) + \mathcal{O}(h^2). \end{aligned}$$

Substituting into the approximation into Eq. (2.6) and integrating yields the scheme

$$u_{n+1} = a_n + h^{-1}L^{-2}(e^{Lh} - 1 - hL)(N(a_n, t_n + h) - N_n),$$

and this is referred to as ETD2RK, which first computes the intermediate solution a_n and then uses it to compute the solution at the next time-step u_{n+1}.

The third order Runge-Kutta form is also derived in similar way by first the intermediate solution a_n and b_n at $t_n + h/2$ and $t_n + h$ respectively using $s = 1$ in (2.9). The next step is to approximate the nonlinear function under the integral in (2.6) using quadratic interpolation polynomial through the points t_n, $t_n + h/2$ and $t_n + h$. This leads to Cox and Matthews' ETD3RK as

$$a_n = e^{Lh/2}u_n + L^{-1}(e^{Lh/2} - 1)N_n, \tag{2.11}$$

$$b_n = e^{Lh}u_n + L^{-1}(e^{Lh} - 1)(2N(a_n, t_n + 1/2) - N_n), \tag{2.12}$$

$$\begin{aligned} u_{n+1} = e^{hL}u_n &+ h^{-2}L^{-3}[-4 - hL + e^{hL}(4 - 3hL + h^2L^2)]N_n + \\ &4h^{-2}L^{-3}[2 + hL + e^{hL}(-2 + hL)]N(a_n, t_n + h/2) + \\ &h^{-2}L^{-3}[-4 - 3hL - h^2L^2 + e^{hL}(4 - hL)]N(b_n, t_n + h) \end{aligned} \tag{2.13}$$

The most used exponential integrator is the Cox and Matthews' fourth-order scheme ETD4RK (Montanelli and Trefethen [2017], Kassam and Trefethen [2005]). The intermediate solutions a_n and b_n approximate the solution at $t_n + h/2$ and a third parameter c_n as the intermediate solution at $t_n + h$. Again, a quadratic interpolation polynomial is used to approximate the nonlinearity. The Cox and Matthews [2002] fourth order scheme is given by

$$a_n = e^{hL/2}u_n + L^{-1}(e^{hL/2} - 1)N_n, \tag{2.14}$$

$$b_n = e^{hL/2}u_n + L^{-1}(e^{hL/2} - 1)N(a_n, t_n + h/2), \tag{2.15}$$

$$c_n = e^{hL/2}a_n + L^{-1}(e^{hL/2} - 1)(2N(b_n, t_n + h/2) - N_n), \tag{2.16}$$

$$\begin{aligned} u_{n+1} = {} & e^{hL}u_n + h^{-2}L^{-3}[-4 - hL + e^{hL}(4 - 3hL + h^2L^2)]N_n + \\ & 2h^{-2}L^{-3}[2 + hL + e^{hL}(-2 + hL)](N(a_n, t_n + h/2) + N(b_n, t_n + h/2)) + \\ & h^{-2}L^{-3}[-4 - 3hL - h^2L^2 + e^{hL}(4 - hL)]N(c_n, t_n + h). \end{aligned} \tag{2.17}$$

The fourth-order scheme is the main time-stepping numerical scheme used in this work.

The Eq. (2.17) does not have the approximate solution at $u(t_{n+1})$ expressed explicitly in terms of the φ functions. A form of Eq. (2.17) given by Krogstad [2005] can be expressed in the Butcher-like tableau

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2}\varphi_1 & & & \\ \frac{1}{2} & 0 & \frac{1}{2}\varphi_1 & & \\ \frac{1}{2} & \frac{1}{2}\varphi_1 & 0 & 0 & \\ \frac{1}{2} & -\frac{1}{2}\varphi_1 & 0 & \varphi_1 & \\ \hline & 4\varphi_3 - 3\varphi_2 + \varphi_1 & -4\varphi_3 + 2\varphi_2 & -4\varphi_3 + 2\varphi_2 & 4\varphi_3 - \varphi_2 \end{array}$$

where $\varphi_k = \varphi_k(hL)$ for $k \geq 1$.

The solution at the internal stages remains the same while the approximate solution at t_{n+1} becomes

$$\begin{aligned}
u_{n+1} = e^{hL}u_n &+ h\left([4\varphi_3(hL) - 3\varphi_2(hL) + \varphi_1(hL)]N(u_n, t_n)\right) \\
&+ h\left([2\varphi_2(hL) - 4\varphi_3(hL)](N(a_n, t_n + h/2) + N(b_n, t_n + h/2))\right) \\
&+ h\left([4\varphi_3(hL) - \varphi_2(hL)]N(c_n, t_n + h)\right). \qquad (2.18)
\end{aligned}$$

2.2.5 Exponential General Linear Methods

The Runge-Kutta form of ETD methods fall under the exponential general linear methods, where given starting values $u_0, u_1, \ldots, u_{q-1}$ at time $t = 0, h, \ldots, (q-1)h$, the approximate solution at next time step is given by (Minchev and Wright [2005], Ostermann et al. [2006])

$$u_{n+1} = e^{hL}u_n + h\sum_{i=1}^{s} B_i(hL)N(v_i) + h\sum_{i=1}^{q-1} V_i(hL)N(u_{n-i}), \qquad (2.19)$$

with q steps u_{n-1} and s stages v_i with $v_1 = u_n$ and

$$v_i = e^{C_i hL}u_n + h\sum_{ji=1}^{i-1} A_{i,j}(hL)N(v_j) + h\sum_{j=1}^{q-1} U_{i,j}(hL)N(u_{n-j}), \quad 2 \le i \le s. \qquad (2.20)$$

The coefficient A, B, C, U and V determines the particular scheme and we represent the exponential general linear method (2.19) in tableau form as

$$\begin{array}{c|cccc|ccc}
c_2 & A_{2,1} & & & & U_{2,1} & \cdots & U_{2,q-1} \\
\vdots & \vdots & \ddots & & & \vdots & & \vdots \\
c_s & A_{s,1} & \cdots & A_{s,s-1} & & U_{s,1} & \cdots & U_{s,q-1} \\
\hline
 & B_1 & \cdots & B_{s-1} & B_s & V_1 & \cdots & V_{q-1}
\end{array}$$

The ETD Adams–Bashford takes the form for $s = 1$ (Ostermann et al. [2006]) in

(2.19) as

$$u_{n+1} = e^{hL}u_n + hB_1(hL)N(u_n) + h\sum_{i=1}^{q-1} V_i(hL)N(u_{n-i}). \tag{2.21}$$

Another form of exponential integrators is the standard integrating factor (IF) method first introduced by Lawson [1967]. The idea makes use of the change of variable

$$v(t) = e^{-Lt}u(t), \tag{2.22}$$

and multiplying Eq. (2.4) by the integrating factor e^{-Lt} and substituting the derivative $v(t)$ leads

$$v_t = e^{-Lt}N(e^{Lt}v, t), \qquad v(0) = u_0. \tag{2.23}$$

The stiff linear part L is not in an explicit form in the differential equation in terms of v in Eq. (2.23) and thus we get rid of the step size or stability constraints. However, the scheme becomes less accurate when a slow-varying nonlinear term is combined with a fast decaying term from the exponential of $-Lt$. The transformed system is solved with a time-stepping scheme such as those of Euler or Runge–Kutta and the approximate solution is then transformed back to the original variable u.

The IF method was generalized by (Krogstad [2005]) to derive the generalized Lawson schemes using the change of variables

$$v(t) = e^{-Lt}u(t) - e^{-Lt}\sum_{j=1}^{q} t^l \varphi_l(Lt)p_{l-1}, \tag{2.24}$$

where p_l are the coefficients of the polynomial interpolant $P(t)$ through the set of points $\{t_{n-l}, N(u_{n-l}, t_{n-l})\}_{l=1}^{q}$. Taking the derivative and substituting into Eq. (2.4), we get the equation in the new variable v as

$$v_t = e^{-Lt}\left(N\left(e^{Lt}v + \sum_{j=1}^{q} t^l\varphi_l(Lt)p_{l-1}, t\right) - P(t)\right), \qquad v(0) = u_0, \tag{2.25}$$

which can be solved with classical fourth order Runge–Kutta.

2.3 Spectral Methods and Polynomial Approximation Theory

Spectral methods are important techniques in numerical analysis and scientific computing. They are methods to discretize the solution of differential and integral equations with orthogonal polynomials or Fourier series.

Spectral methods are global methods, that is, the computations (or derivatives) at any point in the space depend on the information at all other points in the entire domain (Hesthaven et al. [2007]). This is in contrast to the case of finite difference (or elements) termed as local methods whereby the computations at any point depend only on the nearby grid points. In this case, the solution and its derivatives are approximated by a local polynomial interpolant.

The convergence rates of spectral methods assimilate the regularity of the solution: the smoother the function, the faster the convergence rate. This stands in contrast to finite difference or element methods where the convergence rate is also controlled by the mesh (Lui [2012]).

However, traditional spectral methods suffer from a few drawbacks. They can be challenging to implement for complicated domains or complex geometries. Also the method assumes solutions to be sufficiently smooth, hence for nonsmooth functions, it becomes less accurate. This is usually encountered in problems involving shocks and discontinuities.

The basis functions are usually smooth functions and the choice depends on the nature of the solution. For smooth periodic functions with periodic boundary conditions, the natural choice is trigonometric (Fourier) functions and this leads to the so-called Fourier spectral method. For non-periodic functions with non-periodic boundary conditions, one resorts to the use of non-periodic basis functions, mainly polynomial basis functions. In this work, the spectral method is based on the use of polynomials bases which are eigenfunctions of Sturm–Liouville eigenvalue problems. These polynomials are elegant choice of basis due to their nice convergence properties. Examples of these polynomials include the classical orthogonal polynomials of Jacobi, Laguerre, and Hermite.

Polynomial Approximation Theory

The classical orthogonal polynomials are known to be well-suited for expansion of a square integrable function on a bounded interval and the truncated expansions are the best polynomial approximations in their respective Hilbert space (Riesz [1923], Hesthaven et al. [2007], Trefethen [2013]).

The sequence of orthogonal polynomials $\{p_n(x)\}_{n=0}^{\infty}$ are orthogonal with respect to weight function $w(x)$ on some interval (a, b) as

$$\langle p_m, p_n \rangle = \int_a^b p_m(x)p_n(x)w(x)dx = \begin{cases} h_n, & m = n \\ 0, & m \neq n. \end{cases} \tag{2.26}$$

When $h_n = 1$, the polynomials p_n are orthonormal.

The classical orthogonal polynomials are characterized in many ways, including:

1. the polynomials satisfy the Sturm–Liouville differential equation $\sigma p_n'' + \tau p_n' = \lambda_n p_n$, where $\deg(\sigma) \leq 2$, $\deg(\tau) \leq 1$, and $\lambda_n = \frac{n}{2}\left[(n-1)\sigma'' + 2\tau'\right]$. (Bochner [1929], Krall [1941]). Note that Krall extended the result of Bochner to the case where $\langle \cdot, \cdot \rangle$ is a quasi-definite inner product, which is applicable to the Bessel polynomials with quasi–orthogonality in the complex plane.

2. they are the only orthogonal polynomials whose derivatives are also orthogonal polynomials. This makes them suitable for developing spectral methods for solving differential equations (Hahn [1935]).

The most widely used orthogonal polynomials are the Jacobi polynomials (including the special cases of ultraspherical, Chebyshev and Legendre polynomials), Hermite polynomials and Laguerre polynomials.

The orthogonal polynomials also satisfy the three-term recurrence relation

$$p_{n+1}(x) = (A_n x + B_n) p_n(x) - C_n p_{n-1}(x), \tag{2.27}$$

with initial values $p_0(x) = 1$ and $p_1(x) = A_0 x + B_0$ ([DLMF, Eq. 18.19.1]). We summarize properties of the classical orthogonal polynomials in Table 2.1 including the interval of orthogonality, weight and h_n given the standard normalization ([Abramowitz and Stegun, 1972]). In Table 2.1, $\mathcal{A}_n$ denotes $2^{\alpha+\beta+1}\Gamma(n+\alpha+1)\Gamma(n+\beta+1)/((2n+\alpha+\beta+1)\Gamma(n+\alpha+\beta+1)n!)$

As mentioned above, the main idea of spectral methods is to expand any smooth function $u(x)$ in the form

$$u(x) = \sum_{k=0}^{\infty} u_k p_k(x), \tag{2.28}$$

with coefficients defined as

$$u_k = \frac{1}{h_n} \int_a^b p_k(x) u(x) w(x) dx. \tag{2.29}$$

A numerical quadrature can be used to evaluate the integrals and this results in $\mathcal{O}(n^2)$ operation for truncated series with first n terms. A more efficient way is to compute the coefficients using the discrete sine and cosine transforms (DSTs and DCTs) in $\mathcal{O}(n \log n)$ for Chebyshev polynomials of the first, second, third, and fourth kinds.

For any compact interval $[a, b]$, an affine map may be used to expand in Jacobi polynomials. If one endpoint is infinite, then Laguerre polynomials are used.

The convergence of spectral methods depends on the smoothness of the function and subsequent decay of the expansion coefficients as shown in the following theorem

$p_n(x)$	Notation	Interval	$w(x)$	h_n	σ	τ
Jacobi	$P_n^{(\alpha,\beta)}$	$(-1,1)$	$(1-x)^\alpha(1+x)^\beta$	$\mathcal{A}_n$	$1-x^2$	$\beta-\alpha-(\alpha+\beta+2)x$
Ultraspherical	$C_n^{(\lambda)}$	$(-1,1)$	$(1-x^2)^{\lambda-1/2}$	$\frac{2^{1-2\lambda}\pi\Gamma(n+2\lambda)}{(n+\lambda)(\Gamma(\lambda))^2 n!}$	$1-x^2$	$-(2\lambda+1)x$
Legendre	P_n	$(-1,1)$	1	$\frac{2}{2n+1}$	$1-x^2$	$-2x$
Chebyshev (1st Kind)	T_n	$(-1,1)$	$(1-x^2)^{-1/2}$	$\begin{cases} \frac{\pi}{2}, & n>0 \\ \pi, & n=0 \end{cases}$	$1-x^2$	$-x$
Chebyshev (2nd Kind)	U_n	$(-1,1)$	$(1-x)^{1/2}$	$\frac{\pi}{2}$	$1-x^2$	$-3x$
Chebyshev (3rd Kind)	V_n	$(-1,1)$	$(1-x)^{1/2}(1+x)^{-1/2}$	π	$1-x^2$	$-3x$
Chebyshev (4th Kind)	W_n	$(-1,1)$	$(1-x)^{-1/2}(1+x)^{1/2}$	π	$1-x^2$	x
Laguerre	$L_n^{(\alpha)}$	$(0,\infty)$	$e^{-x}x^\alpha$	$\frac{\Gamma(n+\alpha+1)}{n!}$	x	$\alpha+1-x$
Hermite	H_n	$(-\infty,\infty)$	e^{-x^2}	$\pi^{\frac{1}{2}}2^n n!$	1	$-2x$

Table 2.1: Classical Orthogonal Polynomials

(Hesthaven et al. [2007], Lui [2012]).

Theorem 2.4. (Lui [2012]) Let $u \in H_w^{2m}([-1,1])$ for some $m \geq 1$. Suppose that (2.28) holds where p_n is the solution of the Sturm-Liouville eigenvalue problem. Then there exists a constant $0 < c < \infty$ such that

$$|u_k| \leq \frac{c\|u\|_{H_w^{2m}([-1,1])}}{\lambda_k^m}.$$

Theorem 2.5. (Hesthaven et al. [2007]) Let $w(x) = (1-x^2)^{\lambda-1/2}$. For any $u \in H_w^m([-1,1])$, $m \geq 0$. Then for every $N > 0$ there exists a constant $0 < C < \infty$ independent of N such that

$$\|u - \mathcal{P}_N u\|_{L_w^2([-1,1])} \leq CN^{-m}\|u\|_{H_w^m([-1,1])}, \tag{2.30}$$

where $\mathcal{P}_N$ is the canonical orthogonal projection onto the N-dimensional subspace $\text{span}\{C_0^{(\lambda)}, \ldots, C_{N-1}^{(\lambda)}\}$.

Therefore we have spectral convergence for L^2 error of truncation error and this depends on the smoothness of the function u. If $u \in C^\infty([-1,1])$, then the coefficients decay faster than any negative power of N. For an analytic function, the convergence becomes exponential convergence.

2.4 The Ultraspherical Spectral Method

In the ultraspherical spectral method, we consider the solution of a linear differential equation on $[-1,1]$ of the form

$$\Big(a_N\mathcal{D}^N + \cdots + a_1\mathcal{D} + a_0\Big)\,u = f, \quad N \geq 0, \tag{2.31}$$

with the boundary conditions $\mathcal{B}u = c$, which ensure a unique solution of (2.31); and also $a_N \neq 0$ on $[-1, 1]$. In this case the functions f and $a_0, \ldots, a_N$ are so smooth on $[-1, 1]$ that they are well-approximated by degree-m Chebyshev polynomial approximants.

The idea in this spectral method is to represent the solution of (2.31) in the Chebyshev basis and then compute the vector of Chebyshev coefficients of the expansion of the solution. The method seeks to find an infinite vector $\boldsymbol{u} = (u_0, u_1, \ldots)^\top$ such that

$$u(x) = \sum_{k=0}^{\infty} u_k T_k, \quad x \in [-1, 1], \tag{2.32}$$

where T_k is Chebyshev polynomial (of the first kind) of degree k.

Collocation methods based on this representation result in dense linear systems (Trefethen [2000]) and hence the cost can be computationally prohibitive. The ultraspherical spectral method makes use of the relationship between Chebyshev polynomials and derivatives resulting in discretizations which are almost banded.

First Order Differential Equation

Following the idea of [Olver and Townsend, 2013], we consider the first-order linear differential equation to explain the method and the general case will be discussed later. We look for solution of the first-order ODE

$$u'(x) + a(x)u(x) = f(x) \quad \text{and} \quad u(-1) = c, \tag{2.33}$$

where $a(x)$ and $f(x)$ are continuous functions and of bound variation, which ensures the unique representation as uniformly convergent Chebyshev expansion (Trefethen [2013]). Thus we can represent the continuous function $f(x)$ as

$$f(x) = \sum_{k=0}^{\infty} f_k T_k(x). \tag{2.34}$$

We seek the solution of (2.33) in as expansion of Chebyshev polynomials and the main task in this method is solve for the coefficients in this series. The idea is to able to represent the differentiation, conversion from one space to another and the multiplication by variable coefficients as sparse operators.

The first derivative of Chebyshev polynomial of the first kind can be expressed in terms of that of the second kind as

$$\mathcal{D}T_k = \begin{cases} 0 & k = 0, \\ kC_{k-1}^{(1)} & k \geq 1, \end{cases} \tag{2.35}$$

where $C_{k-1}^{(1)} \equiv U_{k-1}$ is Chebyshev polynomial of the second kind of degree $k-1$.

Taking the derivative of the solution (2.32) using Eq. (2.35) scales the Chebyshev coefficients and subsequently changing the basis to $C^{(1)}$ space. That is,

$$u'(x) = \sum_{k=0}^{\infty} ku_k C_{k-1}^{(1)}(x). \tag{2.36}$$

In matrix form, given a vector of Chebyshev coefficients $\boldsymbol{u}$, the derivative in $C^{(1)}$ series is given by $\mathcal{D}_0\boldsymbol{u}$ where $\mathcal{D}_0$ is the differentiation operator

$$\mathcal{D}_0 = \begin{pmatrix} 0 & 1 & & & & \\ & & 2 & & & \\ & & & 3 & & \\ & & & & 4 & \\ & & & & & \ddots \end{pmatrix}$$

The resulting differentiation matrix $\mathcal{D}_0$ is sparse as compared to the dense differentiation matrix from the spectral collocation method. The sparsity of the matrix has a number of advantages, including saving computer memory to fast computations.

For variable coefficients like $a(x)u(x)$ in Eq. (2.33), we have multiplication of two

Chebyshev series. We can represent $a(x)$ and $u(x)$ in Chebyshev expansion as

$$a(x) = \sum_{j=0}^{\infty} a_j T_j(x), \quad \text{and} \quad u(x) = \sum_{k=0}^{\infty} u_k T_k(x),$$

and the product of $a(x)$ and $u(x)$ in Chebyshev series is given as

$$a(x)u(x) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} a_j u_k T_j(x) T_k(x) = \sum_{k=0}^{\infty} c_k T_k(x).$$

We seek to find the vector of coefficients $\boldsymbol{c} = (c_0, c_1, \ldots,)^T$ in terms of a_j and u_k for $j, k = 0, 1, \ldots,$

An explicit formula for the coefficients of products of two Chebyshev expansion was given by Baszenski and Tasche [1997] as

$$c_k = \begin{cases} a_0 u_0 + \frac{1}{2} \sum_{l=1}^{\infty} a_l u_l & k = 0, \\ \frac{1}{2} \sum_{l=0}^{k-1} a_{k-l} u_l + a_0 u_k + \dfrac{1}{2} \sum_{l=1}^{\infty} a_l u_{l+k} + \dfrac{1}{2} \sum_{l=0}^{\infty} a_{l+k} u_l & k \geq 1. \end{cases} \tag{2.37}$$

We can define an operator $\mathcal{M}_0[a]$ as the multiplication of Chebyshev coefficients $\boldsymbol{u} = (u_0, u_1, \ldots,)^\top$ by continuous function function $a(x)$ in Chebyshev series and we can get $\boldsymbol{c} = \mathcal{M}_0[a]\boldsymbol{u}$, where $\mathcal{M}_0[a]$ can be expressed as a sum of Toeplitz matrix and Hankel-plus-rank-1 matrix as

$$\mathcal{M}_0[a] = \frac{1}{2} \left[\begin{pmatrix} 2a_0 & a_1 & a_2 & a_3 & \cdots \\ a_1 & 2a_0 & a_1 & a_2 & \ddots \\ a_2 & 2a_1 & 2a_0 & a_1 & \ddots \\ a_3 & a_2 & a_2 & 2a_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ a_1 & a_2 & a_3 & a_4 & \iddots \\ a_2 & a_3 & a_4 & a_5 & \iddots \\ a_3 & a_4 & a_5 & a_6 & \iddots \\ \vdots & \iddots & \iddots & \iddots & \iddots \end{pmatrix} \right] \tag{2.38}$$

This multiplication operator $\mathcal{M}_0[a]$ is dense. For $a(x)$ being continuous and of bounded variation, its finite Chebyshev expansion converges to $a(x)$ (Trefethen [2013]). Thus, given $\epsilon > 0$, there exists an $m \in \mathbb{N}$ such that

$$\left\| a(x) - \sum_{k=0}^{m-1} a_k T_k(x) \right\|_{L^\infty([-1,1])} < \epsilon.$$

In practice, we can find the coefficients $a_0, \ldots, a_{m-1}$ using the Fast Cosine Transform and subsequently leads to an $m \times m$ dense multiplication operator $\mathcal{M}_0[a]$. However, for $n > m$, the $n \times n$ principal part of $\mathcal{M}_0[a]$ is banded with bandwidth m and also for $a(x)$ smooth, m can be very small. That is, the multiplication operator remains sparse with small bandwidth.

The differentiation operator $\mathcal{D}_0$ takes coefficients in $T_k(x)$ and returns coefficients in $C_k^{(1)}$, in contrast to the multiplication operator $\mathcal{M}_0[a]$ returning coefficients in $T_k(x)$. Adding coefficients in different bases does not make sense. We therefore need to convert one of the coefficients into the other's basis or space, that is, an operator that maps coefficients in Chebyshev basis to $C_k^{(1)}$ space.

Using the relationship between Chebyshev polynomials $T_k(x)$ and $C_k^{(1)}$ by the recurrence relation

$$T_k = \begin{cases} C_0^{(1)} & k = 0, \\ \frac{1}{2} C_1^{(1)} & k = 1, \\ \frac{1}{2}(C_k^{(1)} - C_{k-1}^{(1)}) & k \geq 2, \end{cases} \tag{2.39}$$

we can expand the solution (2.32) using (2.39) as

$$u(x) = \sum_{k=0}^{\infty} u_k T_k(x) = u_0 C^{(1)}(x) + \frac{1}{2} u_1 C^{(1)}(x) + \frac{1}{2} \sum_{k=2}^{\infty} u_k \Big(C_k^{(1)}(x) - C_{k-2}^{(1)}(x) \Big). \tag{2.40}$$

Writing few terms from the summation and rearranging gives $u(x)$ as

$$u(x) = \Big(u_0 - \frac{1}{2} u_2 \Big) C_0^{(1)}(x) + \sum_{k=1}^{\infty} \frac{1}{2} (u_k - u_{k+2}) C_k^{(1)}(x). \tag{2.41}$$

Thus, given a vector of Chebyshev coefficients $\boldsymbol{u}$, a conversion operator $\mathcal{S}_0$ converts them to coefficients in $C_k^{(1)}$ as $\mathcal{S}_0\boldsymbol{u}$ where the conversion operator is given by

$$\mathcal{S}_0 = \begin{pmatrix} 1 & & -\frac{1}{2} & & & \\ & \frac{1}{2} & & -\frac{1}{2} & & \\ & & \frac{1}{2} & & -\frac{1}{2} & \\ & & & \ddots & & \ddots \end{pmatrix}.$$

Once again we have the conversion operator being sparse and banded. We represent function $f(x)$ in Eq. (2.33) in Chebyshev series and subsequently convert the coefficients to that $C^{(1)}$ basis. We represent the differential equation (2.33) as

$$(\mathcal{D}_0 + \mathcal{S}_0\mathcal{M}_0[a])\boldsymbol{u} = \mathcal{S}_0\boldsymbol{f}, \tag{2.42}$$

which can also be written as

$$\mathcal{L}\boldsymbol{u} = \mathcal{S}_0\boldsymbol{f}, \tag{2.43}$$

where $\mathcal{L}$ is defined as $\mathcal{L} = \mathcal{D}_0 + \mathcal{S}_0\mathcal{M}_0[a]$, and $\boldsymbol{f}$ and $\boldsymbol{u}$ are the vectors of coefficients of expansion $f(x)$ and $u(x)$ respectively.

In practice, we truncate the infinite dimensional operators to work with the finite dimensional space. We define an $n \times \infty$ projection operator which maps an infinite dimensional operator in $\mathbb{C}^\infty$ to a finite case $\mathbb{C}^n$ as

$$\mathcal{P}_n = (I_n, \mathbf{0}), \tag{2.44}$$

where I_n is an $n \times n$ identity matrix. For example, applying the projection operator on the differentiation operator as $\mathcal{P}_n\mathcal{D}_0\mathcal{P}_n^\top$ truncates the the rows and columns to $n \times n$ matrix with the last rows being zeros. Hence the structure of $\mathcal{P}_n\mathcal{L}\mathcal{P}_n^\top$ makes it easy to impose the boundary conditions by replacing the last row with the evaluations at the boundary. One can permute this last row to be the first row so that the resulting linear system is close to an upper triangular matrix. Thus, we approximate the solution of (2.33) by first solving for the coefficients $u_0, u_1, \ldots, u_{n-1}$ in the linear system

$$\begin{pmatrix} T_0(-1) & T_1(-1) & \cdots & T_{n-1}(-1) \\ & & \mathcal{P}_{n-1} L \mathcal{P}_n^\top & \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} c \\ (\mathcal{P}_{n-1} \mathcal{S}_0 \mathcal{P}_n^\top)(\mathcal{P}_n \boldsymbol{f}) \end{pmatrix}, \tag{2.45}$$

and the we get the solution as

$$\tilde{u}(x) = \sum_{k=0}^{n-1} u_k T_k(x).$$

Olver and Townsend described an adaptive QR algorithm to solve the linear system in $\mathcal{O}(m^2 n)$ operations, where m is the bandwidth of the banded discretization of $\mathcal{L}$. The complexity of the *factorization* is $\mathcal{O}(m^2 n)$. However, once the linear operator is factorized, solving linear systems then costs only $\mathcal{O}(mn)$.

Higher Order Differentiation Equations

Consider the N^{th}-order linear ODE (2.31) with K boundary conditions $\mathcal{B}u = \boldsymbol{c} \in \mathbb{C}^K$.

Representing the solution of (2.31) in Chebyshev expansion as we did in the first-order case (2.32), we would need the three operators: differentiation, multiplication and conversion to solve (2.31).

We use the ultraspherical polynomial $C_k^{(\lambda)}$ of degree k for $\lambda = 1, 2, \ldots$. This family of orthogonal polynomials satisfies the following relation to compute the derivatives

$$\frac{dC_k^{(\lambda)}}{dx} = \begin{cases} 0 & k = 0, \\ 2\lambda C_{k-1}^{(\lambda+1)} & k \geq 1. \end{cases} \tag{2.46}$$

Given the solution $u(x)$ in Chebyshev expansion and using the derivative relations (2.35), we have for arbitrary $\lambda \geq 1$

$$\frac{d^\lambda u(x)}{dx^\lambda} = \sum_{k=1}^{\infty} k u_k \frac{d^{\lambda-1} C_{k-1}^{(1)}(x)}{dx^{\lambda-1}},$$

and subsequently taking the derivative $\lambda - 1$ times using relation (2.46), we have

$$\frac{d^\lambda u(x)}{dx^\lambda} = 2^{\lambda-1}(\lambda-1)! \sum_{k=\lambda}^{\infty} k u_k C_{k-\lambda}^{(\lambda)}(x)$$

In matrix form, we have a sparse representation representation as the λth-order differentiation operator $\mathcal{D}_\lambda$ as

$$\mathcal{D}_\lambda = 2^{\lambda-1}(\lambda-1) \begin{pmatrix} \overbrace{0 \quad \cdots \quad 0}^{\lambda \quad \text{times}} & \lambda & & & \\ & & \lambda+1 & & \\ & & & \lambda+2 & \\ & & & & \ddots \end{pmatrix}$$

The operator $\mathcal{D}_\lambda$ maps the vector coefficients of the Chebyshev expansion to vector coefficients of the expansion in $C^{(\lambda)}$.

Since $\mathcal{D}_\lambda$ return a vector of coefficients in $C^{(\lambda)}$, we would need conversion operators to convert of coefficients resulting from lower order derivatives. We use the following relationship between ultraspherical polynomials

$$C_k^{(\lambda)} = \begin{cases} C_0^{(\lambda+1)} & k=0, \\ \frac{\lambda}{\lambda+1} C_1^{(\lambda+1)} & k=1, \\ \frac{\lambda}{\lambda+k}(C_k^{(\lambda+1)} - C_{k-2}^{(\lambda+1)}) & k \geq 2. \end{cases} \tag{2.47}$$

The general case using Eq. (2.47) for $\lambda \geq 1$, where $\mathcal{S}_\lambda$ converts coefficients in $C^{(\lambda)}$ to that of $C^{(\lambda+1)}$ series as

$$\mathcal{S}_\lambda = \begin{pmatrix} 1 & 0 & -\frac{\lambda}{\lambda+2} & & \\ & \frac{\lambda}{\lambda+1} & 0 & -\frac{\lambda}{\lambda+3} & \\ & & \frac{\lambda}{\lambda+2} & 0 & \ddots \\ & & & \frac{\lambda}{\lambda+3} & \ddots \\ & & & & \ddots \end{pmatrix}.$$

We need one more operator for the multiplication of function $a(x)u(x)$, particularly to represent the product of two ultraspherical expansions. We expand the function $a(x)$ and $u(x)$ in $C^{(\lambda)}$ series as

$$a(x) = \sum_{j=0}^{\infty} a_j C_j^{(\lambda)}(x) \qquad \text{and} \qquad u(x) = \sum_{k=0}^{\infty} u_k C_k^{(\lambda)}(x)$$

we have the product as

$$a(x)u(x) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} a_j u_k C_j^{(\lambda)}(x) C_k^{(\lambda)}(x). \tag{2.48}$$

Using of Carlitz [1961] approach on product of two ultraspherical polynomials as

$$C_j^{(\lambda)}(x) C_k^{(\lambda)}(x) = \sum_{s=0}^{min(j,k)} c_s^\lambda(j,k) C_{j+k-2s}^{(\lambda)}(x), \tag{2.49}$$

where

$$c_s^\lambda(j,k) = \frac{j+k+\lambda-2s}{j+k+\lambda-s}\frac{(\lambda)_s(\lambda)_{j-s}(\lambda)_{k-s}}{s!(j-s)!(k-s)!}\frac{(2\lambda)_{j+k-s}}{(\lambda)_{j+k-s}}\frac{(j+k-2s)!}{(2\lambda)_{j+k-2s}}, \tag{2.50}$$

and

$$(\lambda)_k = \frac{\Gamma(\lambda+k)}{\Gamma(\lambda)} = \frac{(\lambda+k-1)!}{(\lambda-1)!}$$

is termed as the Pochhammer symbol.

Substituting Eq. (2.49) into Eq. (2.48) and rearranging the summation yields

$$a(x)u(x) = \sum_{j=0}^{\infty}\Bigg(\sum_{k=0}^{\infty}\sum_{s=\max(0,k-j)}^{k} a_{2s+j-k}c_s^k(k,2s+j-k)u_k\Bigg)C_j^{(\lambda)}(x). \tag{2.51}$$

Thus in matrix representation form, we have the (j,k) entry of the operator representing the multiplication of $a(x)$ in $C^{(\lambda)}$ expansion is given by

$$\mathcal{M}_\lambda[a]_{j,k} = \sum_{s=\max(0,k-j)}^{k} a_{2s+j-k}c_s^k(k,2s+j-k), \qquad j,k\geq 0. \tag{2.52}$$

As mentioned previous case, the function $a(x)$ can be approximated finite summation of first m coefficients

$$a(x) = \sum_{j=0}^{m-1} a_j C_j^{(\lambda)}(x), \tag{2.53}$$

and the resulting truncated operator $\mathcal{P}_n\mathcal{M}_\lambda[a]\mathcal{P}_n^\top$ is banded with bandwidth m. One can also approximate the $a(x)$ by finite number of Chebyshev coefficients and then convert it to $C^{(\lambda)}$ space using the truncated conversion operators.

Having obtained all the operators required solve (2.31), the discretized ODE becomes

$$\boldsymbol{\mathcal{L}u} = \mathcal{S}_{N-1}\cdots\mathcal{S}_0\boldsymbol{f}, \tag{2.54}$$

where

$$\mathcal{L} = \mathcal{M}_N[a_N]\mathcal{D}_N + \sum_{\lambda=1}^{N-1} \mathcal{S}_{N-1} \cdots \mathcal{S}_\lambda \mathcal{M}[a_\lambda]\mathcal{D}_\lambda + \mathcal{S}_{N-1} \cdots \mathcal{S}_0 \mathcal{M}_0[a_0].$$

The K boundary conditions are imposed by replacing the last K row of the truncated part of the differential operator in (2.54). We can permute this to have the first K rows being the boundary condition to make the resulting system close to a upper triangular matrix. Thus we solve for coefficients $u_0, \dots u_{n-1}$ in the following linear system

$$\begin{pmatrix} \mathcal{B}\mathcal{P}_n^\top \\ \mathcal{P}_{n-K} L \mathcal{P}_n^\top \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{c} \\ \mathcal{P}_{n-K}\mathcal{S}_{N-1} \dots . \mathcal{S}_0 \boldsymbol{f} \end{pmatrix}.$$

The approximate solution of (2.31) is then given by

$$\tilde{u}(x) = \sum_{k=0}^{n-1} u_k T_k(x).$$

Although ultraspherical spectral methods are known to be efficient, the representation of the solution and subsequently using boundary bordering leads to a non-symmetric matrix even for self-adjoint differential operators (Olver and Townsend [2013]). This lack of symmetry leads to inaccuracy in the computations of the eigenvalues especially at the high modes for an eigenvalue differential problem.

2.5 Symmetrizing The Ultraspherical Spectral Method

For self-adjoint eigenvalue problems, the ultraspherical representation results in a non-symmetric matrix and thus leading to complex spectrum. The method proposed by Aurentz and Slevinsky [2020] seeks to symmetrize the ultraspherical spectral methods using basis recombination of orthogonal polynomials.

Following the idea presented in their paper, we consider the self-adjoint Sturm–Liouville problem

$$\mathcal{L}u := -\mathcal{D}^2 u = \lambda u, \qquad u(\pm 1) = 0. \tag{2.55}$$

The main idea is to represent the solution in a basis that satisfies the boundary conditions.

If we expand the solution of (2.55) as weighted normalized ultraspherical polynomials

$$u(x) = \sum_{k=0}^{\infty} u_n (1-x^2)\tilde{C}_n^{(\frac{3}{2})}(x),$$

then the negative second-order differentiation of u is a diagonal matrix with entries $d_n = (n+1)(n+2)$, when the basis for the range is $\{\tilde{C}_n^{(\frac{3}{2})}(x)\}_{n=0}^{\infty}$. Thus, the multiplication operator of $1-x^2$, which is a symmetric pentadiagonal matrix, represents expansions in $(1-x^2)\tilde{C}_n^{(\frac{3}{2})}(x)$ in the unweighted basis $\tilde{C}_n^{(\frac{3}{2})}(x)$, and is given by (Aurentz and Slevinsky [2020])

$$\mathcal{M}[1-x^2] = \begin{pmatrix} a_0 & 0 & b_0 & \\ 0 & a_1 & & \ddots \\ b_0 & & a_2 & \\ & \ddots & & \ddots \end{pmatrix},$$

where

$$a_n = \frac{2(n+1)(n+2)}{(2n+1)(2n+5)}, \qquad \text{and} \qquad b_n = -\sqrt{\frac{(n+1)(n+2)(n+3)(n+4)}{(2n+3)(2n+5)^2(2n+7)}}.$$

The representation leads to

$$D\boldsymbol{u} = \lambda M\boldsymbol{u}, \tag{2.56}$$

where where D is a diagonal and M is symmetric pentadiagonal and positive definite. Thus the discretization leads to the generalized symmetric-definite pencil (D, M) where the generalized eigenvalues are real.

2.5.1 Basis Recombination

Considering the classical problem (1.2) with homogeneous boundary condition and with the Hilbert space H and the constrained Hilbert space $H_{\mathcal{B}}$ defined in section 1.1.1. Let $\{\phi_n\}_{n=0}^{\infty}$ be an orthonormal polynomials basis for H with $\deg(\phi_n) = n$.

For the linearly independent boundary conditions in $\mathcal{B}$, let $\{\rho_n\}_{n=0}^{\infty}$ be the basis recombination of the orthonormal polynomials $\{\phi_n\}_{n=0}^{\infty}$ so that $\mathcal{B}\rho_n = 0$. Then the model problem above suggests the existence of a conversion operator A which is lower triangular and banded such that

$$\begin{pmatrix} \rho_0 & \rho_1 & \rho_2 & \cdots \end{pmatrix} = \begin{pmatrix} \phi_0 & \phi_1 & \phi_2 & \cdots \end{pmatrix} A. \tag{2.57}$$

Using the QR factorization of $A = QR$, the unitary matrix Q maps the orthonormal polynomial basis for H to the orthonormal basis $\{\psi_n\}_{n=0}^{\infty}$ for the constrained space $H_{\mathcal{B}}$

$$\begin{pmatrix} \psi_0 & \psi_1 & \psi_2 & \cdots \end{pmatrix} = \begin{pmatrix} \phi_0 & \phi_1 & \phi_2 & \cdots \end{pmatrix} Q, \tag{2.58}$$

and we also get the operator R from

$$\begin{pmatrix} \rho_0 & \rho_1 & \rho_2 & \cdots \end{pmatrix} = \begin{pmatrix} \psi_0 & \psi_1 & \psi_2 & \cdots \end{pmatrix} R. \tag{2.59}$$

By defining another set of polynomials $\{\sigma_n\}_{n=0}^{\infty}$ such that

$$\begin{pmatrix}\sigma_0 & \sigma_1 & \sigma_2 & \cdots\end{pmatrix} R^* = \begin{pmatrix}\psi_0 & \psi_1 & \psi_2 & \cdots\end{pmatrix}, \tag{2.60}$$

and since

$$\begin{pmatrix}\psi_0 & \psi_1 & \psi_2 & \cdots\end{pmatrix} Q^* = \begin{pmatrix}\phi_0 & \phi_1 & \phi_2 & \cdots\end{pmatrix}. \tag{2.61}$$

Using $A^* = R^* Q^*$, we get

$$\begin{pmatrix}\sigma_0 & \sigma_1 & \sigma_2 & \cdots\end{pmatrix} A^* = \begin{pmatrix}\phi_0 & \phi_1 & \phi_2 & \cdots\end{pmatrix}. \tag{2.62}$$

As mentioned above, for the problem Eq. (1.4) the choice for ϕ_n is the normalized Legendre polynomials $\phi_n(x) = \tilde{P}_n(x)$ and the recombination $\rho_n(x) = (1-x^2)\tilde{P}_n^{(1,1)}(x)$ and the orthonormal basis for $H_{\mathcal{B}}$ is $\rho_n(x) = (1-x^2)\tilde{P}_n^{(2,2)}(x)$. Using these four bases, [Aurentz and Slevinsky, 2020] showed that resulting Petrov–Galerkin for the problem is symmetric-definite and banded where the solution is represented as $u(x) = \boldsymbol{\rho}^\top \boldsymbol{v} = \boldsymbol{\psi}^\top R\boldsymbol{v} = \boldsymbol{\psi}^\top \boldsymbol{u}$.

The solution can written in an expansion of the basis $\boldsymbol{\psi}$ and a vector of coefficients $\boldsymbol{u} = (u_0, u_1, u_2, \cdots)^\top$ as

$$u(x) = \sum_{n=0}^{\infty} u_n \psi_n(x) = \boldsymbol{\psi}^\top \boldsymbol{u}.$$

2.5.2 The Petrov-Galerkin is banded and self-adjoint

Using the orthonormal polynomial basis ψ_n for $H_{\mathcal{B}}$, it follows that that the Ritz–Galerkin scheme is self-adjoint:

$$\boldsymbol{\mathcal{L}\psi}^\top \boldsymbol{u} = \boldsymbol{\psi}^\top L\boldsymbol{u} = \lambda w\boldsymbol{\psi}^\top \boldsymbol{u} = \boldsymbol{\psi}^\top \lambda M\boldsymbol{u},$$

where $L = L^*$ and $M = M^*$ is positive-definite. Using $\boldsymbol{\rho}^\top = \boldsymbol{\psi}^\top R$ in Eq. (2.59) we have

$$\boldsymbol{\psi}^\top LR\boldsymbol{v} = \boldsymbol{\psi}^\top \lambda MR\boldsymbol{v},$$

and using $\boldsymbol{\sigma}^\top R^* = \boldsymbol{\psi}^\top$ in Eq. (2.60),

$$\boldsymbol{\sigma}^\top R^* LR\boldsymbol{v} = \lambda \boldsymbol{\sigma}^\top R^* R\boldsymbol{v}.$$

we have that R^*LR is self-adjoint and R^*MR is self-adjoint and positive-definite.

To show bandedness, we have

$$\boldsymbol{\mathcal{L}}\boldsymbol{\rho}^\top \boldsymbol{v} = \lambda w \boldsymbol{\rho}^\top \boldsymbol{v},$$

and using $\boldsymbol{\rho}^\top = \boldsymbol{\phi}^\top A$ in Eq. (2.57), we have $\boldsymbol{\mathcal{L}}\boldsymbol{\rho}^\top \boldsymbol{v} = \boldsymbol{\mathcal{L}}\boldsymbol{\phi}^\top A\boldsymbol{v}$. Suppose L_B is banded operator representing the discretization using the ultrapherical spectral method, then there exists an upper triangular and banded above conversion operator so that $\boldsymbol{\mathcal{L}}\boldsymbol{\phi}^\top = \boldsymbol{\phi}^\top C^{-1} L_B$ and using $\boldsymbol{\phi}^\top = \boldsymbol{\sigma}^\top A^*$ in Eq. (2.62), we get

$$\boldsymbol{\mathcal{L}}\boldsymbol{\phi}^\top A\boldsymbol{v} = \boldsymbol{\phi}^\top C^{-1} L_B A\boldsymbol{v} = \boldsymbol{\sigma}^\top A^* C^{-1} L_B A\boldsymbol{v} = \lambda \boldsymbol{\sigma}^\top A^* M_B A\boldsymbol{v}.$$

The bandedness comes from the fact that $L_B A$ is banded below and C^{-1} is an upper triangular matrix which will not extend the bandwidth the $L_B A$ any lower. Since A^* is also upper-triangular and the Petrov-Galerkin scheme is self-adjoint, then the scheme is banded.

2.5.3 Bound on the Projection Error

Let $\mathcal{P}_n : H_{\mathcal{B}} \to H_{\mathcal{B}}$ be the canonical orthogonal projection onto the n-dimensional subspace span$\{\psi_0, \ldots, \psi_{n-1}\}$ of the recombined basis and the discrete form as $P_n : \ell^2 \to \ell^2$. Using the solution $u(x) = \boldsymbol{\rho}^\top \boldsymbol{v} = \boldsymbol{\psi}^\top R\boldsymbol{v} = \boldsymbol{\psi}^\top \boldsymbol{u} \in H_{\mathcal{B}}$, Aurentz and Slevinsky [2020] showed that

$$\|u - \mathcal{P}_n u\|_{H_{\mathcal{B}}} \leq 2\|R\|_{\ell^2 \to \ell^2} \|\boldsymbol{v} - P_n \boldsymbol{v}\|_{\ell^2}. \tag{2.63}$$

This implies that the error committed by projecting u onto the recombined space n-dimensional subspace $\{\psi_0, \ldots, \psi_{n-1}\}$ is bounded by a constant multiple of the discrete discrete truncation error $\boldsymbol{v} - P_n \boldsymbol{v}$

The bound on the projection error in H can also be obtained using the canonical projection $\mathcal{P}_n : H \to H$ onto the span$\{\phi_0, \ldots, \phi_{n-1}\}$ and using $u(x) = \boldsymbol{\rho}^\top \boldsymbol{v} = \boldsymbol{\phi}^\top A\boldsymbol{v} = \boldsymbol{\phi}^\top \boldsymbol{w} \in H$. In this case, Aurentz and Slevinsky [2020] also showed that

$$\|u - \mathcal{P}_{n+2N} u\|_H \leq 2\|A\|_{\ell^2 \to \ell^2} \|\boldsymbol{v} - P_n \boldsymbol{v}\|_{\ell^2}. \tag{2.64}$$

3

Eigen-Decomposition Approach

3.1 Introduction

The implementation of exponential integrators encounters the difficulty in computing the φ functions, which are the coefficients g_m in (2.9). For ETD1, we require the computation of

$$\varphi_1(z) = \frac{e^z - 1}{z}. \tag{3.1}$$

This "definition" (3.1) suggests $z = 0$ is not in the domain of φ_1, yet it is (based on the Maclaurin series in Eq. (3.3)). This functions suffers from round-off errors due to cancellation errors for z close to the origin. It is a well known computational problem in numerical analysis (Higham [2002]). The limiting form of $\varphi_1(z)$ as $z \to 0$ is 1 but the direct computation leads to inaccurate results and this problem is worse in higher forms of exponential integrators.

The general form of the φ functions is given by Hochbruck and Ostermann [2005]

$$\varphi_l(z) = \frac{\varphi_{l-1}(z) - \dfrac{1}{l!}}{z} = \frac{e^z - \displaystyle\sum_{j=0}^{l-1} \frac{z^j}{j!}}{z^l}, \qquad \text{for} \qquad l \geq 1, \tag{3.2}$$

where $\varphi_0(z) = e^z$. For example, we have

$$\varphi_2(z) = \frac{e^z - 1 - z}{z^2}, \qquad \text{and} \qquad \varphi_3(z) = \frac{e^z - 1 - z - \frac{z^2}{2}}{z^3},$$

which appear in second and third order ETD respectively. Also we have $\varphi_l(0) = 1/l!$.

A simple approach to alleviate this problem of computation is to use the Maclaurin series expansion for values of z close to the origin (Cox and Matthews [2002]). The Maclaurin series representation of the φ functions is given by

$$\varphi_l(z) = \sum_{k=0}^{\infty} \frac{z^k}{(k+l)!}. \tag{3.3}$$

It follows that the φ functions are entire, since the Maclaurin series has infinite radius of convergence. Therefore, the Maclaurin series can be evaluated with matrix argument. However subtractive cancellation in the summation is the cause of numerical instability which leads inaccurate results especially if the spectrum is in the left-half plane. Therefore, our strategy utilizes both the Maclaurin series and the asymptotic formula for $\varphi_l(z)$ in Eq. (3.2) for pointwise computation after a (generalized) eigendecomposition of the matrices reveals the spectrum of the linear operators.

3.2 Eigendecomposition with Homogeneous Boundary Conditions

The method of Aurentz and Slevinsky [2020] synthesizes a symmetric-definite and banded discretization to represent a stiff self-adjoint linear differential operator with polynomial coefficients. Discretizing and truncating so that $A, B, L \in \mathbb{R}^{n\times n}$, we have

$$L = B^{-1}A, \tag{3.4}$$

to represent the linear differential operator in Eq. (1.1) and as elaborated in Eqs. (1.2) and (1.3). It follows from a congruence transformation with a Cholesky factorization of B that the generalized spectral decomposition of L exists (Van Loan and Golub [1983]), the spectrum is real, and the generalized eigenvectors can be chosen to be real and B-orthogonal. Let $\Lambda \in \mathbb{R}^{n\times n}$ denote the diagonal matrix of generalized eigenvalues and $V \in \mathbb{R}^{n\times n}$ the corresponding eigenvectors. It follows that for any $\lambda \in \mathbb{R}$,

$$V^\top (A - \lambda B)V = \Lambda - \lambda I.$$

For any entire function $f : \mathbb{C} \to \mathbb{C}$, the corresponding matrix function $f(L)$ may be defined in terms of the eigendecomposition of $L \in \mathbb{R}^{n\times n}$ (provided it exists) by

$$f(L) = V f(\Lambda)V^{-1}, \quad \text{where} \quad L = V\Lambda V^{-1},$$

and where the diagonal matrix function acts entry-wise on the main diagonal. If $L = B^{-1}A$ as above, then the B-orthogonality of the generalized eigenvectors circumvents the necessity for the inverse, since $V^{-1} = V^\top B$.

Since the φ functions are all entire, it follows that

$$\varphi_l(L) = V\varphi_l(\Lambda)V^\top B.$$

Therefore computing the φ functions on matrices is simply done by evaluating the functions on the diagonal entries of a diagonal matrix provided the spectral decomposition exists.

3.3 Non-Homogeneous Problems

We consider a self-adjoint non-homogeneous boundary condition, $\mathcal{B}u = \boldsymbol{c} \neq 0$, where we represent the solution as

$$u(x,t) = v(x,t) + w(x). \tag{3.5}$$

where $v(x,t)$ satisfies the homogeneous boundary condition, $\mathcal{B}v = 0$, and $w(x)$ satisfies the non-homogeneous boundary condition, $\mathcal{B}w = \boldsymbol{c}$. For a time dependent boundary condition such as $\mathcal{B}u = \boldsymbol{c}(t)$, then w will also be time dependent as $w(x,t)$. We intend to solve the problem in v using the approach for a homogeneous case and the main idea in this section is the formulation of $w(x)$ using basis recombination based on the given boundary condition.

3.3.1 Basis Recombination

We would like to show that $\{\rho_n(x)\}_{n=0}^{\infty}$ is a basis for the complete Hilbert space $H = L^2([-1,1])$. By the Stone–Weierstrass approximation theorem and density of $C([-1,1])$ in $L^2([-1,1])$, it is known that any degree-graded polynomial sequence is a basis for $H = L^2([-1,1])$. The recombined polynomials, however, are only degree-graded in the generalized sense that there exists a positive integer ν such that $\deg(\rho_{n+\nu}) > \deg(\rho_n)$. Therefore, it is not immediately obvious that they are a basis for H. There are a few equivalent proofs, including: a linear-algebraic proof that the annihilator in Eq. (2.57) has dense column space; and, a functional-analytic proof that recombinations are dense in $H_{\mathcal{B}}$ which is itself dense in H. Both of these proofs use standard constructions from their area, but do not illustrate the central approximation-theoretic issue with the use of the recombinations as a basis.

Instead, we will show directly for the case of Dirichlet boundary conditions that the original orthonormal polynomials have 2-norm-convergent representations in the recombined basis that also converge pointwise almost everywhere but not uniformly. Thus, recombinations give rise to a Gibbs-like phenomenon (Wilbraham [1848]), when approximating functions in $H \setminus H_{\mathcal{B}}$. To overcome the slow decay in the coefficients of the expansions in recombined polynomials, we augment the recombinations with a finite set of low-degree polynomials to strip off the inhomogeneity at the boundary.

Let $\phi_n(x) = \tilde{P}_n(x)$ be the normalized Legendre polynomials. Let $\mathcal{B}u = u(\pm 1) = 0$ denote Dirichlet boundary conditions. Let $\rho_n(x) = \tilde{P}_n(x) - \sqrt{\dfrac{2n+1}{2n+5}}\tilde{P}_{n+2}(x)$ be the recombined normalized Legendre polynomials so that $\mathcal{B}\rho_n \equiv 0$.

Theorem 3.1. Let k be a non-negative integer and let $u(x) = \boldsymbol{\rho}^\top \boldsymbol{v}$, where:

$$v_n = \begin{cases} \sqrt{\dfrac{2k+1}{2n+1}} & k \leq n,\ k+n \text{ even}, \\ 0 & \text{otherwise.} \end{cases}$$

Then the canonical orthogonal projections:

1. converge 2-normwise to the normalized Legendre polynomials of degree k:

$$\lim_{N\to\infty} \|\tilde{P}_k - \mathcal{P}_{N+k}u\| = 0,$$

2. and for every $x \in (-1, 1)$:

$$\lim_{N\to\infty} \left|\tilde{P}_k(x) - \mathcal{P}_{N+k}u(x)\right| = 0.$$

Proof. We assume without loss of generality that N is odd.

1. From the annihilator:

$$A = \begin{pmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ -\sqrt{1/5} & 0 & 1 & & & \\ & -\sqrt{3/7} & 0 & 1 & & \\ & & -\sqrt{5/9} & 0 & 1 & \\ & & & \ddots & \ddots & \ddots \end{pmatrix},$$

it follows that the coefficients v_n are the infinite-dimensional solution to the linear system $A\boldsymbol{v} = \boldsymbol{e_k}$, where $\boldsymbol{e_k}$ is the k^{th} canonical basis vector.

For the norm of the approximation error, we convert the projections in the recombined basis to the orthonormal basis. Using the fact that:

$$AP_{N+k}\boldsymbol{v} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ v_k \\ v_{k+1} \\ \vdots \\ v_n - \sqrt{\dfrac{2n-3}{2n+1}}v_{n-2} \\ \vdots \\ -\sqrt{\dfrac{2N+2k-1}{2N+2k+3}}v_{N+k-1} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ -\sqrt{\dfrac{2k+1}{2N+2k+3}} \end{pmatrix},$$

we find:

$$\lim_{N\to\infty} \|\tilde{P}_k - \mathcal{P}_{N+k}u\| = \lim_{N\to\infty} \left\| \tilde{P}_k - \tilde{P}_k + \sqrt{\frac{2k+1}{2N+2k+3}} \tilde{P}_{N+k+1} \right\|,$$
$$= \lim_{N\to\infty} \sqrt{\frac{2k+1}{2N+2k+3}} = 0.$$

2. For the pointwise convergence, the recombined expansion is a telescoping series:

$$\mathcal{P}_{N+k}u(x) = \sqrt{2k+1} \sum_{n=k,2}^{N+k-1} \left(\sqrt{\frac{1}{2n+1}} \tilde{P}_n(x) - \sqrt{\frac{1}{2n+5}} \tilde{P}_{n+2}(x) \right).$$

Since:

$$\lim_{N\to\infty} \left| \tilde{P}_k(x) - \mathcal{P}_{N+k}u(x) \right| = \sqrt{k+1/2} \lim_{N\to\infty} |P_{N+k+1}(x)|,$$

the result follows by Bernstein's inequality ([DLMF, Eq. 18.14.7]):

$$|P_n(x)| < \sqrt{\frac{4}{\pi(2n+1)}} \frac{1}{(1-x^2)^{\frac{1}{4}}}.$$

□

As can be expected by the slow decay in the truncated coefficients, the practical effectiveness of the recombined polynomials as a basis for H is questionable. Therefore, when approximating functions $f \in H \setminus H_{\mathcal{B}}$, we use a *redundant* function set $\operatorname{span}\{1, x\} \cup \{\rho_n(x)\}_{n=0}^{\infty}$. The Gibbs-like phenomenon is shown in Figure 3.1 for approximations of $\tilde{P}_0(x)$ and $\tilde{P}_1(x)$ in the recombined basis ρ_n. Next, we discuss how to deduce the precise augmentation from the boundary conditions.

The boundary condition can be represented as an infinite-dimensional matrix-vector product (Aurentz and Slevinsky [2020]) with the matrix $B \in \mathbb{C}^{2N\times\infty}$, where $\mathcal{B}u = \mathcal{B}\boldsymbol{\phi}^\top \boldsymbol{w} = B\boldsymbol{w}$.

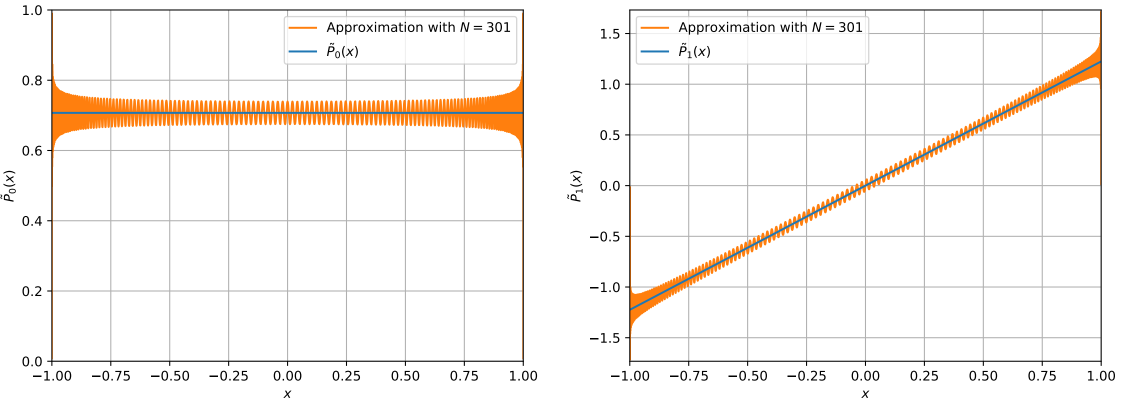


Figure 3.1: The polynomial approximations of $\tilde{P}_0(x)$ (*left*) and $\tilde{P}_1(x)$ (*right*) in the recombined basis.

Considering the linear system

$$\begin{pmatrix} b_{0,0} & b_{0,1} & b_{0,2} & \cdots \\ b_{1,0} & b_{1,1} & b_{1,2} & \cdots \\ \vdots & \vdots & \vdots & \\ b_{2N-1,0} & b_{2N-1,1} & b_{2N-1,2} & \cdots \end{pmatrix} \begin{pmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,2N-1} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,2N-1} \\ w_{2,0} & w_{2,1} & \cdots & w_{2,2N-1} \\ \vdots & \vdots & & \vdots \end{pmatrix} = I, \tag{3.6}$$

then a solution, if it exists, would provide $2N$ polynomials $w_0, \ldots, w_{2N-1}$, such that

$$\mathcal{B}\left(c_0 w_0 + c_1 w_1 + \cdots + c_{2N-1} w_{2N-1}\right) = \boldsymbol{c}.$$

3.3.2 Dirichlet Case

Considering a second-order problem with nonhomogeneous boundary conditions $\mathcal{B}u = \boldsymbol{c}$, we look for bases $u_L(x)$ and $u_R(x)$ satisfying $\mathcal{B}u_L = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathcal{B}u_R = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

We pose the boundary conditions on Legendre polynomials P_n of degree n such that

$$\mathcal{B}\begin{pmatrix} u_L & u_R \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$$

For Dirichlet boundary conditions $\mathcal{B}u = u(\pm 1) = \boldsymbol{c}$ and using $P_n(\pm 1) = (\pm 1)^n$, then Eq. (3.6) becomes

$$\begin{pmatrix} 1 & -1 & 1 & -1 & \cdots \\ 1 & 1 & 1 & 1 & \cdots \end{pmatrix} \begin{pmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \\ w_{20} & w_{21} \\ w_{30} & w_{31} \\ \vdots & \vdots \end{pmatrix} = I. \tag{3.7}$$

We take the first two columns of the evaluations as matrix, say $\hat{B} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ and left multiplication of Eq. (3.7) by $\hat{B}^{-1}$ gives

$$\begin{pmatrix} 1 & 0 & 1 & 0 & \cdots \\ 0 & 1 & 0 & 1 & \cdots \end{pmatrix} \begin{pmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \\ w_{20} & w_{21} \\ w_{30} & w_{31} \\ \vdots & \vdots \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}. \tag{3.8}$$

This shows that there are a countably infinite finite-dimensional set of polynomials that can be used to remove the boundary components. We use the lowest degree polynomials out of convenience.

Taking a finite dimension, we get the 'left' basis

$$u_L(x) = w_{00}P_0 - w_{10}P_1 = \frac{1}{2}P_0 - \frac{1}{2}P_1 = \frac{1}{2} - \frac{1}{2}x,$$

such that $u_L(-1) = 1$, $u_L(1) = 0$, and similarly for the 'right' basis

$$u_R(x) = w_{00}P_0 + w_{10}P_1 = \frac{1}{2}P_0 + \frac{1}{2}P_1 = \frac{1}{2} + \frac{1}{2}x,$$

with $u_R(-1) = 0$, $u_R(1) = 1$. Hence the recombination of these bases can be used to formulate $w(x)$ for non-homogeneous Dirichlet boundary condition. Thus the second order problem with Dirichlet boundary $u(-1) = c_1$ and $u(1) = c_2$, we have $w(x)$ given by

$$w(x) = c_1 u_L(x) + c_2 u_R(x) = c_1\left(\frac{1}{2}P_0 - \frac{1}{2}P_1\right) + c_2\left(\frac{1}{2}P_0 + \frac{1}{2}P_1\right),$$

which clearly satisfies the non-homogeneous boundary condition.

3.3.3 Neumann Case

Neumann boundary conditions, $\mathcal{B}u = u'(\pm 1) = \mathbf{c}$, illustrate a subtle issue. From $P_n'(\pm 1) = (\pm 1)^{n+1}\dfrac{n(n+1)}{2}$, then Eq. (3.6) becomes

$$\begin{pmatrix} 0 & 1 & -3 & 6 & \cdots \\ 0 & 1 & 3 & 6 & \cdots \end{pmatrix} \begin{pmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \\ w_{20} & w_{21} \\ w_{30} & w_{31} \\ \vdots & \vdots \end{pmatrix} = I. \tag{3.9}$$

Taking the first two columns of the evaluations includes a column of zeros and hence the linear system becomes rank deficient. A similar problem was encountered by Aurentz and Slevinsky [2020] where they suggested doubling the columns and subsequently using the Moore–Penrose pseudoinverse to solve the linear system. We could also add a single column instead of doubling. Thus we take the first three columns of the evaluations as a matrix

$$\hat{B} = \begin{pmatrix} 0 & 1 & -3 \\ 0 & 1 & 3 \end{pmatrix}.$$

Since $\hat{B}$ is rectangular, we left multiply Eq. (3.9) by its Moore–Penrose pseudoinverse given by

$$\hat{B}^{+} = \begin{pmatrix} 0 & 0 \\ 1/2 & 1/2 \\ -1/6 & 1/6 \end{pmatrix},$$

and choosing the lowest degree polynomials to satisfy the boundary conditions, we get

$$\begin{aligned} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \\ w_{20} & w_{21} \end{pmatrix} &= \hat{B}^{+}, \\ \begin{pmatrix} 0 & 0 \\ w_{10} & w_{11} \\ w_{20} & w_{21} \end{pmatrix} &= \begin{pmatrix} 0 & 0 \\ 1/2 & 1/2 \\ -1/6 & 1/6 \end{pmatrix}. \end{aligned}$$

The following quadratic bases can be used for the recombination

$$u_L(x) = w_{10}P_1 - w_{20}P_2 = \frac{1}{2}P_1 - \frac{1}{6}P_2,$$

$u_L'(-1) = 1$, $u_L'(1) = 0$, and also

$$u_R(x) = w_{11}P_1 - w_{21}P_2 = \frac{1}{2}P_1 - \frac{1}{6}P_2,$$

with $u'_R(-1) = 0$, $u'_R(1) = 1$.

The formulation of $w(x)$ for non-homogeneous Neumann boundary condition follows the idea in the previous case. For Neumann boundary condition $u'(-1) = c_1$ and $u'(1) = c_2$ for the second order problem, we have

$$\begin{aligned} w(x,t) =& c_1 u_L(x) + c_2 u_R(x), \\ =& c_1\Big(\frac{1}{2}P_1 - \frac{1}{6}P_2\Big) + c_2\Big(\frac{1}{2}P_1 + \frac{1}{6}P_2\Big). \end{aligned}$$

For a simple problem with negative and positive slopes at the end points, that is, $u'(-1) = -1$ and $u'(1) = 1$, we have $w(x,t) = 1/2x^2 - 1/6$, which satisfies the nonhomogeneous boundary condition.

4

Rational Approximation

4.1 Introduction

In Chapter 3, we used a generalized eigendecomposition to assist in the evaluation of the φ functions of matrices. Alternative methods are based on rational approximations. The most well-known method to generate rational approximations to matrix functions is to reformulate via the Cauchy integral formula and to discretize by a quadrature rule. Another more direct approach is the approximation of the φ function via the method of Carathéodory–Fejér.

The computation of matrix functions by the Cauchy integral formula is a powerful tool in scientific computing. By matrix function, we mean a scalar function f such as $\exp(z)$, $\log(x)$, or $z^{1/2}$ and a matrix $A \in \mathbb{C}^{n\times n}$ such that $f(A)$ is of the same dimension as A. The use of contour integrals to compute matrix functions was studied by Talbot [1979] to invert Laplace transforms. Several authors have used this idea for solution of PDEs where the integral is approximated by means of trapezoidal rule which converges exponentially (Trefethen and Weideman [2014]). Matrix functions via contour integrals can be represented in the Cauchy integral formula, which is also known as Dunford (Dunford-Taylor) formula (Higham [2008], Hale et al. [2008]). For an analytic function f and a square matrix A, we have

$$f(A) = \frac{1}{2\pi i}\int_{\Gamma} f(z)(zI - A)^{-1}dz, \tag{4.1}$$

where Γ is a positively-oriented contour in the region of analyticity of f and winding once around the spectrum of A. The resulting quadrature formula for the integral can be associated with rational approximation which will be discussed in Section 4.4 for functions of matrices.

4.2 The Cauchy Integral for φ functions

Kassam and Trefethen [2005] first introduced the idea of contour integral to compute the $\varphi_l(t)$ by considering the integral representation on a contour Γ enclosing t. In a simple case, using a circle of unit radius centred at t and approximating the integral with trapezoidal rule, the Cauchy integral becomes

$$\varphi_l(t) = \frac{1}{2\pi i}\int_{\Gamma} \frac{\varphi_l(z)}{z - t}dz \approx \frac{1}{n}\sum_{k=1}^{n} \varphi_l(t + e^{i\theta_k}), \tag{4.2}$$

where $\theta_k = 2\pi k/n$ are the n points on the circle, parameterized by angle. One has to ensure that the circle does not get too close to the origin. Examples of such contours are the Talbot-type contours such as parabolæ, hyperbolæand cotangent (Trefethen et al. [2006]), and the particular choice of the contour depends on the problem (Montanelli and Bootland [2016]).

Consider the Cauchy integral representation of the exponential function $\varphi_0(t) = e^t$ on a contour Γ which encloses t in Eq. (4.2).

Let $z : \mathbb{R} \to \Gamma$ be analytic. Then after variable transformation, the integral in

Eq. (4.2) becomes

$$e^t = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{e^{z(\theta)}}{z(\theta) - t} z'(\theta) d\theta. \tag{4.3}$$

The term $e^{z(\theta)}$ leads to exponential decay of the integrand as $|\theta| \to \infty$ provided the contour Γ begins and ends in the left-half plane. Thus there is minimal exponential error committed when one truncates the contour's preimage, $\mathbb{R}$, to a finite interval. Following the idea of Trefethen et al. [2006], we consider the interval $[-\pi, \pi]$ and we take n equispaced point θ_k with step size $2\pi/n$. The integral is then approximated by "trapezoidal" quadrature rule[1] as

$$e^t \approx \frac{-i}{n} \sum_{k=1}^{n} \frac{e^{z(\theta_k)}}{z(\theta_k) - t} z'(\theta_k). \tag{4.4}$$

This method can be used to compute matrix exponential provided we find an elegant choice of contour with optimized parameters enclosing the spectrum of the matrix.

The exponential decay of the integrand is lost if we generalize the approach to evaluate the functions φ_l, $l > 0$, which decay algebraically in the left-half plane. Trefethen and Schmelzer [2007] proposed an idea of additional reparametrization to enforce exponential decay of the integrand for any φ function as summarized in the following theorem.

Theorem 4.1. (Trefethen and Schmelzer [2007]) Let Γ be a closed contour encircling the points 0 and $t \in \mathbb{C}$ with winding number 1. Then

$$\varphi_l(t) = \frac{1}{2\pi i} \int_{\Gamma} \frac{e^z}{z^l} \frac{1}{z - t} dz \tag{4.5}$$

The proof of this theorem is quite technical but essentially based on residue

[1] Here, trapezoidal is in quotations because the first and last contributions are not halved.

theorem.

The integral representation (4.5) can be used to compute the function φ_l without involving the function φ_l in the integrand. Using this idea for the matrix functions, we evaluate the φ functions as

$$\varphi_l(hL) = \frac{1}{2\pi i}\int_\Gamma \frac{e^{hz}}{(hz)^l}(zI - L)^{-1}dz. \tag{4.6}$$

In practice, we are interested in computing the matrix-vector product $\varphi_l(hL)\boldsymbol{v}$ rather than the $\varphi_l(hL)$. That is,

$$\varphi_l(hL)\boldsymbol{v} = \frac{1}{2\pi i}\int_\Gamma \frac{e^{hz}}{(hz)^l}(zI - L)^{-1}\boldsymbol{v}dz. \tag{4.7}$$

Parametrizing a chosen contour as $z(\theta)$, we have

$$\varphi_l(hL)\boldsymbol{v} = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{e^{hz(\theta)}z'(\theta)}{(hz(\theta))^l}(z(\theta)I - L)^{-1}\boldsymbol{v}d\theta. \tag{4.8}$$

Considering the discretized linear operator in this research where $L = B^{-1}A$, we have

$$zI - L = zI - B^{-1}A = B^{-1}(zB - A).$$

We approximate the integral with trapezoidal rule and using n nodes θ_k in a fixed interval $[-\pi, \pi]$ of spacing $2\pi/n$ as

$$\varphi_l(hL)\boldsymbol{v} \approx -\frac{i}{n}\sum_{k=1}^{n} \frac{e^{hz(\theta_k)}z'(\theta_k)}{(hz(\theta_k))^l}[z(\theta_k)B - A]^{-1}B\boldsymbol{v}, \tag{4.9}$$

$$\approx \frac{i}{n}\sum_{k=1}^{n} w_k[z(\theta_k)B - A]^{-1}B\boldsymbol{v}, \tag{4.10}$$

where w_k is given by

$$w_k = -\frac{e^{hz(\theta_k)} z'(\theta_k)}{(hz(\theta_k))^l}.$$

The method (4.10) applied to evaluate any φ_l to a vector solves a banded linear system at each node $z(\theta_k)$ of the trapezoidal rule which can be achieved in $\mathcal{O}(n)$ complexity. Also exploiting symmetry of the contour, we can solve at only half of the nodes and then take twice the real part of the results.

4.3 Choice of Contour and Error Estimates

A key aspect of contour integral approach for matrix functions is the selection of a suitable contour which usually depends on the problem, and subsequently using an efficient numerical integration to approximate the integral on the contour. In the case of this research, the contour must enclose the spectrum of the linear operator L but it cannot get too close to the spectrum otherwise the norm of the resolvent $(zI - L)^{-1}$ grows unbounded.

We consider a hyperbolic contour of the form (Weideman and Trefethen [2007])

$$z = \mu(1 + \sin(i\theta - \alpha)), \quad -\infty < \theta < \infty, \tag{4.11}$$

where $\mu \geq 0$ controls the width of the contour and the parameter α determines the asymptotic angle of the hyperbola. Weideman and Trefethen [2007] have derived optimal values of the parameters for the computation of the Bromwich integral which has the exponential in the integrand. Figure 4.1 shows a schematic diagram of the hyperbola (4.11).

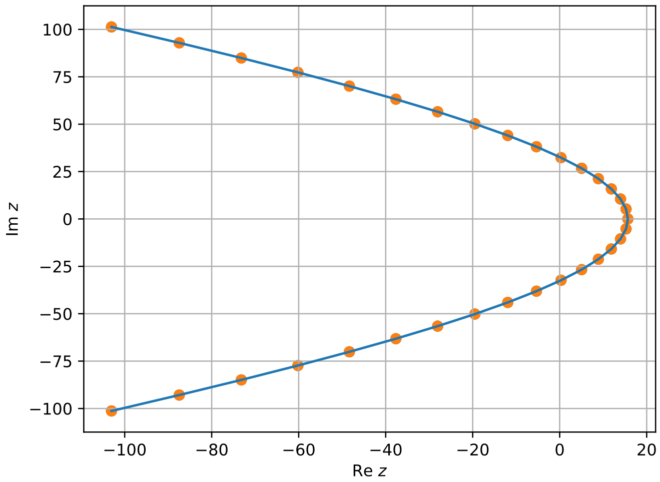


Figure 4.1: Hyperbolic contour that winds around the negative real axis also enclosing the origin.

The error analysis is done by considering the absolutely convergent integral

$$I = \int_{-\infty}^{\infty} g(\theta) d\theta, \tag{4.12}$$

with infinite and finite trapezoidal rule approximations

$$I_h = h \sum_{k=-\infty}^{\infty} g(kh), \qquad I_h^{[n]} = h \sum_{k=-n}^{n} g(kh). \tag{4.13}$$

The error (between I and $I_h^{[n]}$) in approximating the integral results from the discretization error (DE) and the truncation error (TE) given by

$$|I - I_h^{[n]}| \leq |I - I_h| + |I_h - I_h^{[n]}|.$$

and we have the DE and TE as

$$\text{DE} = |I - I_h|, \quad \text{TE} = |I_h - I_h^{[n]}|.$$

To estimate the discretization error, we use an idea from a well-known theorem of Martensen [1968] for the trapezoidal rule.

Consider $w = \theta + iv$ with θ and v real. Suppose that $g(w)$ is analytic in the strip $d < v < c$ for some $c > 0, d > 0$ with $g(w) \to 0$ uniformly as $|w| \to \infty$ in the strip. In this case, we have the function $g(w)$ being a complex valued. Thus we consider the analytic properties of the integrand in both the upper and lower half-planes, leading to different error estimates in each of the half-planes. Suppose that for some $M_+(c) > 0$, $M_-(d) > 0$ the function $g(w)$ satisfies

$$\int_{-\infty}^{\infty} |g(\theta + ic)| d\theta \leq M_+(c), \qquad \int_{-\infty}^{\infty} |g(\theta - id)| d\theta \leq M_-(d).$$

Then the discretization error is bounded by

$$|I - I_h| \leq \mathrm{DE}_+ + \mathrm{DE}_-,$$

where

$$\mathrm{DE}_+ = \frac{M_+(c)}{e^{2\pi c/h} - 1}, \qquad \mathrm{DE}_- = \frac{M_-(d)}{e^{2\pi d/h} - 1}. \tag{4.14}$$

The original idea (Martensen [1968]) considers $g(w)$ to be real-valued function where $c = d$ and $M_- = M_+ = M$ with the error estimate

$$|I - I_h| \leq \frac{2M(c)}{e^{2\pi c/h} - 1}.$$

The convergence rate of the approximation of the integral depends on the width of the strip of analyticity (Weideman and Trefethen [2007], Trefethen and Weideman [2014]).

Estimate for DE_+: Consider the upper half-plane with $c > 0$, we have the line

$$w = \theta + ic, \qquad -\infty < \theta < \infty,$$

which has image under the map (4.11)

$$z = \mu(1 - \sin(\alpha + c)\cosh(\theta)) + i\mu\cos(\alpha + c)\sinh(\theta). \tag{4.15}$$

From the error estimates (4.14), the decay rate of the error is maximized for a large c in the upper half-plane. When c is increased from 0 in Eq. (4.15), the width of the hyperbola gets smaller until it degenerates into the negative real axis when $c = \pi/2 - \alpha$ as

$$z = \mu(1 - \sin(\alpha + c)\cosh(\theta)).$$

However, at the same time, since the spectrum of L is on the negative real axis, $M_+(c)$ becomes unbounded. This effect is due to the norm of the resolvent $(z(w)I - L)^{-1}$ getting large as $z(w)$ as approaches or lies on the negative real axis. Thus c should not be $\pi/2 - \alpha$. One way to resolve is to set the maximum of value of c to be $\pi/2 - \alpha - \epsilon$ for some $0 < \epsilon \ll 1$ to maximize the decay rate of the error.

To see the effect of the resolvent, we consider a diagonalizable $L = V\Lambda V^{-1}$:

$$\begin{aligned}(z(w)I - L)^{-1} &= (z(w)I - V\Lambda V^{-1})^{-1},\\ &= [V(z(w)I - \Lambda)V^{-1}]^{-1},\\ &= V(z(w)I - \Lambda)^{-1}V^{-1}.\end{aligned}$$

The bound of the norm is given by Trefethen and Embree [2005]

$$||(z(w)I - L)^{-1}||_2 \leq \kappa(V)||(z(w)I - \Lambda)^{-1}||_2 = \frac{\kappa(V)}{\operatorname{dist}(z(w), \sigma(L))},$$

where $\kappa(V)$ is the condition number of V, $\sigma(L)$ is the spectrum of L and the term $\operatorname{dist}(z(w), \sigma(L))$ represents the distance of the point $z(w)$ to the spectrum $\sigma(L)$.

Thus we take the error estimate in the upper half-plane as

$$\mathrm{DE}_+ = \mathcal{O}(e^{-2\pi(\pi/2-\alpha-\epsilon)/h}), \quad \text{as} \quad h \to 0. \tag{4.16}$$

Estimate for DE$_-$: We consider the lower half-plane $d > 0$ where the image of the map of $w = \theta - id$ is given by

$$z = \mu(1 + \sin(d - \alpha)\cosh\theta) + i\mu\sinh\theta\cos(d - \alpha). \tag{4.17}$$

According to the estimates from the theorem, we seek to maximize d in order to enhance the exponential decay of the error.

When d is increased from 0, the hyperbola opens up wide until it becomes vertical line in the half-plane when $d = \alpha$, where we have

$$z = i\mu\sinh\theta.$$

Thus we have the liming value of d is when $d = \alpha$. In this case the norm of the resolvent is not a limiting factor since the hyperbola is far off the negative real axis.

Thus we take $d = \alpha$ and also considering the contribution of the term e^z in the integrand leads to the error estimate

$$\mathrm{DE}_- = \mathcal{O}(e^{\mu-2\pi\alpha/h}), \quad \text{as} \quad h \to 0. \tag{4.18}$$

Estimate for TE: The infinite series (4.13) is truncated to a finite summation for implementation purposes and this leads to the truncation error (TE). Suppose that $g(w)$ decays rapidly, one can estimate truncation error using the magnitude of the last term retained in the finite sum in Eq. (4.13). The truncation error is estimated as

$$\begin{aligned} \mathrm{TE} &= \mathcal{O}(g(hn)), && (4.19) \\ &= \mathcal{O}(e^{\mu-\mu\sin\alpha\cosh(hn)}) \qquad \text{as} \qquad n \to \infty. && (4.20) \end{aligned}$$

In our implementation, we consider spacing on the contour as $h = 2\pi/n$ and the the optimal parameter values derived by Weideman and Trefethen [2007] as $\alpha = 1.1721$, $\mu = 4.4921n$. The total error is estimated by combining discretization and truncation error estimates as

$$En = \mathcal{O}(e^{-(\pi-2\alpha-2\epsilon)n/2}) \qquad \text{as} \quad n \to \infty \tag{4.21}$$

4.4 Carathéodory–Fejér Approach

We explore a third approach which is the so-called Carathéodory–Fejér (CF) approximation, to compute the φ function. CF approximation is a near-best rational approximation introduced by Trefethen and Gutknecht [1983] for approximations on a unit disc, which may be transplanted to finite or infinite intervals by a Möbius transformation. It is sometimes regarded as exact in practice due to its high accuracy. The main idea is to find singular values and corresponding vectors of a Hankel matrix of the coefficients of the polynomial (Taylor or Chebyshev) expansion of the function ([Trefethen, 2013]). The underlying concept was studied by Carathéodory, Fejér, Shur and Takagi to estimate the Fourier and Laurent series (Magnus [1994]). The CF method has been used in a number of applications including the solution of

PDEs and Talbot quadrature (Trefethen and Schmelzer [2007]).

4.5 CF Approximation on the negative real line

The use of CF method for computing $\exp(z)$ on the negative real line was introduced by Trefethen et al. [2006], which subsequently led to extending it to evaluate the φ functions of exponential integrators.

The partial fraction expansion of the rational approximations of type (n, n) is given by

$$r_n(z) = \frac{p_n(z)}{q_n(z)} = r_\infty + \sum_{j=1}^{n} \frac{c_j}{z - z_j}, \tag{4.22}$$

where c_j is the residue of the pole z_j and $r_\infty = r(\infty)$. For an entire function that is real-valued on the real line, the polynomial $q_n(z)$ has real coefficients and thus the poles come in conjugate pairs. The poles and residues of the rational approximation (4.22) can be interpreted as the quadrature nodes and weights in the contour integral approach, respectively.

The best rational approximation of $\exp(z)$ on the negative real axis using (4.22) has error decreasing at rate $(9.28903)^{-n}$ ([Trefethen et al., 2006]) as $n \to \infty$. To make this approach more general for all the φ functions, one can use the same poles approximating one of the φ functions to compute all the other functions instead of sampling the poles and residues for each of the functions. In this case, we can use the poles and residues of the approximation of $\exp(z)$ for other φ functions.

Using proposition 4.2 by Trefethen and Schmelzer [2007], suppose we define a matrix B_z as

$$B_z = \begin{pmatrix} z & 1 \\ 0 & 0 \end{pmatrix},$$

where we have $B_z^0 = I$ and also

$$B_z^n = \begin{pmatrix} z^n & z^{n-1} \\ 0 & 0 \end{pmatrix}.$$

From the Taylor form (3.3) of the φ functions, we have

$$\begin{aligned}
\varphi_l(B_z) =& \sum_{k=l}^{\infty} \frac{1}{k!} B_z^{k-l} \\
=& \frac{1}{l!} I + \sum_{k=l+1}^{\infty} \frac{1}{k!} B_z^{k-l} \\
=& \begin{pmatrix} \sum_{k=l}^{\infty} \frac{1}{k!} z^{k-l} & \sum_{k=l+1}^{\infty} \frac{1}{k!} z^{k-l-1} \\ 0 & \frac{1}{l!} \end{pmatrix} \\
=& \begin{pmatrix} \varphi_l(z) & \varphi_{l+1}(z) \\ 0 & \varphi_l(0) \end{pmatrix}
\end{aligned}$$

Suppose we have the rational approximation of $\varphi_l(z)$

$$r_n^{(l)}(z) = r_\infty + \sum_{j=1}^{n} \frac{c_j}{z - z_j},$$

and also the inverse of $(B_z - z_j I)$ given by

$$(B_z - z_j I)^{-1} = \begin{pmatrix} (z - z_j)^{-1} & z^{-1}(z - z_j)^{-1} \\ 0 & -z_j^{-1} \end{pmatrix}.$$

.

From the above identity, we have $(1, 2)$ entry of $\varphi_l(B_z)$ as the function $\varphi_{l+1}(z)$

given $\varphi_l(z)$. Thus we can approximate φ_{l+1} from the $(1,2)$ entry of $r_n^{(l)}(B_z)$ as

$$r_n^{(l+1)}(z) = \sum_{j=1}^{n} \frac{c_j z_j^{-1}}{z - z_j},$$

and this leads to the recurrence relation for general case for computing the φ functions

$$r_n^{(l+k)}(z) = \sum_{j=1}^{n} \frac{c_j z_j^{-k}}{z - z_j}, \quad k \in \mathbb{Z}. \tag{4.23}$$

In particular, suppose we have the poles and residues approximating $\varphi_0(z) = \exp(z)$, then the approximation for the other φ functions is given by

$$\varphi_k(z) \approx \sum_{j=1}^{n} \frac{c_j z_j^{-k}}{z - z_j}, \quad k \geq 0. \tag{4.24}$$

Similarly, one can use the poles and residues of the CF approximation of φ_1 to evaluate the rest of the φ functions following the identity (4.23).

We compute the $\varphi_l(z)$ for $l = 0, 1, 2, 3$ for z close to zero $(1e-16)$. We translate the code for computing the poles and residues of a function on the negative real axis which was originally in MATLAB by [Trefethen et al., 2006] into Julia. Using the poles of rational approximation of φ_0 and φ_1, the absolute error $\varphi_l - r_n^{(l)}$ for degree $n = 12$ rational approximation is shown in Table 4.1

	Using poles of φ_0 approximant	using poles of φ_1 approximant
φ_0	$3.1565861036142 45e-12$	$4.716049772923725e-10$
φ_1	$1.5803403030645313e-10$	$1.3367085216486885e-13$
φ_2	$2.638425922185661e-9$	$6.512956840509787e-12$
φ_3	$1.7679756297850346e-8$	$1.0279468942719916e-10$

Table 4.1: Error committed for approximating the φ_l using the poles of φ_0 and φ_1 generated with the CF approximation on the negative real axis following the identity (4.23)

.

This approximation method (4.24) can be extended to compute the matrix functions where z is a matrix and that is the focus of this research. As mentioned in the contour integral section 4.2, we are interested in the matrix-vector product $\varphi_k(hL)\boldsymbol{v}$ for a vector $\boldsymbol{v}$. Thus from Eq. (4.24), we have

$$\varphi_k(hL)\boldsymbol{v} \approx \sum_{j=1}^{n} c_j z_j^{-k}(hL - z_j I)^{-1}\boldsymbol{v}, \quad k \geq 0.$$

Again, using the discretized linear operator of the form $L = B^{-1}A$, we get the approximation of the matrix-vector product as

$$\varphi_k(hL)\boldsymbol{v} \approx \sum_{j=1}^{n} c_j z_j^{-k}(hA - z_j B)^{-1}B\boldsymbol{v}, \quad k \geq 0. \tag{4.25}$$

The approximation (4.25) solves n linear systems for a type (n, n) rational approximation of the matrix-vector product $\varphi_k(hL)\boldsymbol{v}$ but each linear system can solved in linear complexity for banded matrices A and B. Also since the poles come in conjugate pairs, we can take advantage of the symmetry and solve about half the linear systems by taking twice the results of the real part.

5

Numerical Experiments

5.1 Introduction

In this section, we present the results of numerical experiments on some PDEs with different boundary boundary to investigate the convergence of the time-stepping scheme combined with the structure preserving spectral method of Aurentz and Slevinsky [2020]. We do that in both space and time where solution is simulated from an initial time $t = 0$ up to a final time $t = T$. For temporal convergence, an extremely small time step is used to estimate an exact solution $u^{exact}(x, t = T)$ and compute the relative error between this solution and the solutions $u(x, t = T)$ obtained by relatively larger time steps. For spatial convergence, the 'exact' solution $u^{exact}(x, t = T)$ is obtained using a small time step and a large number of coefficients and we then measure error by varying the number of coefficients of the solution $u(x, t = T)$. At the final time $t = T$, the L^2 relative error is given by

$$\text{Relative Error} = \frac{||u^{exact}(x, t = T) - u(x, t = T)||_2}{||u^{exact}(x, t = T)||_2}$$

We also measure the computer time for the simulations, that is, the execution time (time-stepping) against the error and the pre-computation time against the number of coefficients (of the polynomial expansion of the solution). The implementations were carried out on a 2.4 GHz Intel i5 MacBook Pro with 8 GB of RAM.

5.1.1 Computing the Nonlinearity

At every time step, we store the coefficients of the polynomial expansion of the approximate solution. Thus the computation of the nonlinearity in the discretized form Eq. (2.4) has to be done in the physical or value space. For example, the nonlinearity in the Allen–Cahn equation is $N(u,t) = u - u^3$. There is the need to transform between coefficients in different spaces using fast transforms.

Consider a sufficiently smooth function $f(x)$ in different polynomial expansions as

$$f(x) = \sum_{n=0}^{\infty} a_n \rho_n(x) = \sum_{n=0}^{\infty} b_n P_n(x) = \sum_{n=0}^{\infty} c_n T_n(x),$$

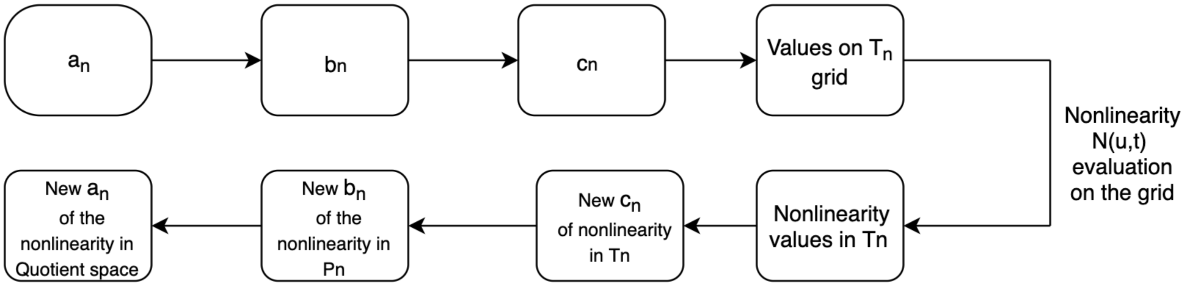
where ρ_n, P_n and T_n are the recombined basis, Legendre polynomials and Chebyshev polynomials, respectively. We have their respective expansion coefficients as a_n, b_n and c_n. Given the coefficients in the quotient space (recombined basis), we compute the given nonlinearity as follows:

- the coefficients a_n in the constrained Hilbert space are first transformed to coefficients b_n of the Legendre polynomial expansion by applying the conversion operator in Eq. (2.57) in linear complexity;

- the Legendre coefficients b_n are then transformed to Chebyshev coefficients c_n using a fast Legendre–Chebyshev transform. One efficient transform is the fast multipole-like method by Alpert and Rokhlin [1991] and it achieves the conversion between Legendre and Chebyshev coefficients with linear complexity. A recent method which is fast and simple was described by Hale and Townsend [2014] based on asymptotic formula of Chebyshev polynomial and has complexity of $\mathcal{O}(N(\log N)^2/\log\log N)$;

- The Chebyshev coefficients are then transformed to values on the Chebyshev grid via the DCT, where the nonlinearity can be evaluated pointwise in value space;
- We now transform the values of the nonlinearity on the grid back to Chebyshev coefficients;
- Using the Chebyshev–Legendre transform, we get the coefficients of the nonlinearity in Legendre series; and,
- To be able to stay in the quotient space, the infinite coefficients in Legendre space are projected to a finite number of coefficients in constrained space.

The above steps are summarized in Figure 5.1



a_n
bn
cn
Values on T_n grid
Nonlinearity N(u,t) evaluation on the grid
Nonlinearity values in Tn
New c_n of nonlinearity in Tn
New b_n of the nonlinearity in Pn
New a_n of the nonlinearity in Quotient space

Figure 5.1: Steps for evaluating the nonlinearity.

5.2 Homogeneous Boundary Conditions

5.2.1 The Allen–Cahn Equation

Experiment using Eigendecomposition

We solve the Allen–Cahn equation with initial condition $u(x, t = 0) = (1-x^2)(0.53x+0.47\sin(-1.5\pi x)-x)$. The solution using ETD4RK up to the final time $T = 64$ confirms the preservation of symmetry and the metastability of the Allen–Cahn equation as shown in the Figure 5.2.

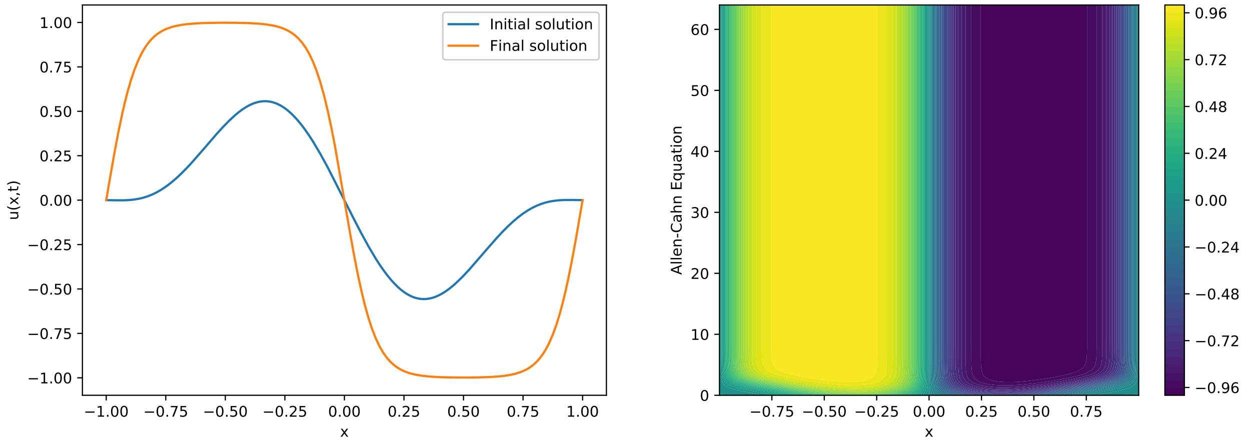


Figure 5.2: Time evolution of the Allen–Cahn equation on $x \in [-1, 1]$.

We show spatial convergence where we compute the relative error between the solution obtained with $N = 300$ coefficients with time step $h = 2^{-8}$ and the solutions obtained by varying the number of coefficients. For temporal convergence, we use time $h = 2^{-8}$ with $N = 300$ coefficients to get the 'exact' solution and then compute the relative norm for solutions obtained by varying the time step with $N = 150$. The plots in the Figure 5.3 shows spectral spatial convergence and fourth order temporal convergence which confirm accuracy of the methods used to solve the problem up to the final time $T = 1$.

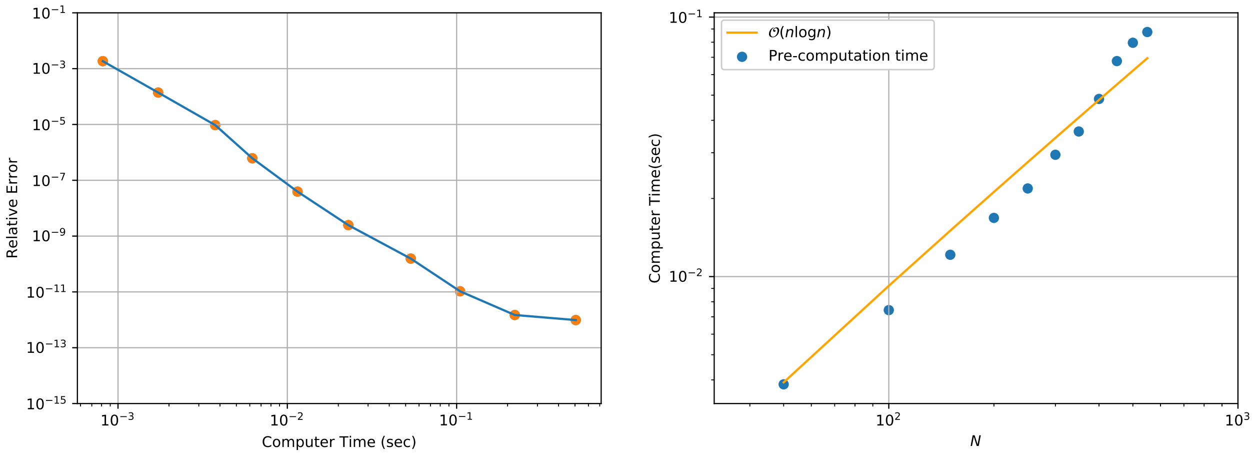


Figure 5.4: Execution time against the relative error (*left*) and pre-computation time against the number of solution coefficients (*right*).

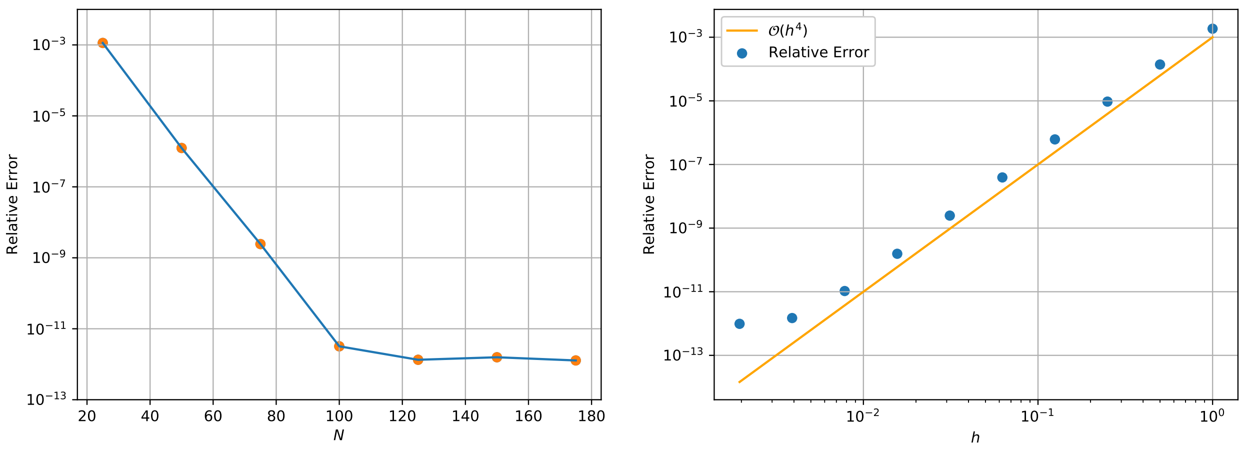


Figure 5.3: Spatial (*left*) and temporal (*right*) convergence at final time $T = 1$.

Experiment using Contour Integrals

We implement the contour integral method to compute the φ functions as discussed in Section 4. We solve the Allen–Cahn equation using Krogstad [2005] form of ETD4RK where the approximate solution at t_{n+1} in Eq. (2.18) is expressed in terms of the φ functions. Each of the φ functions is evaluated on the hyperbolic contour (4.11). For experimental purposes, it is observed that using at least $n = 2^6$ nodes

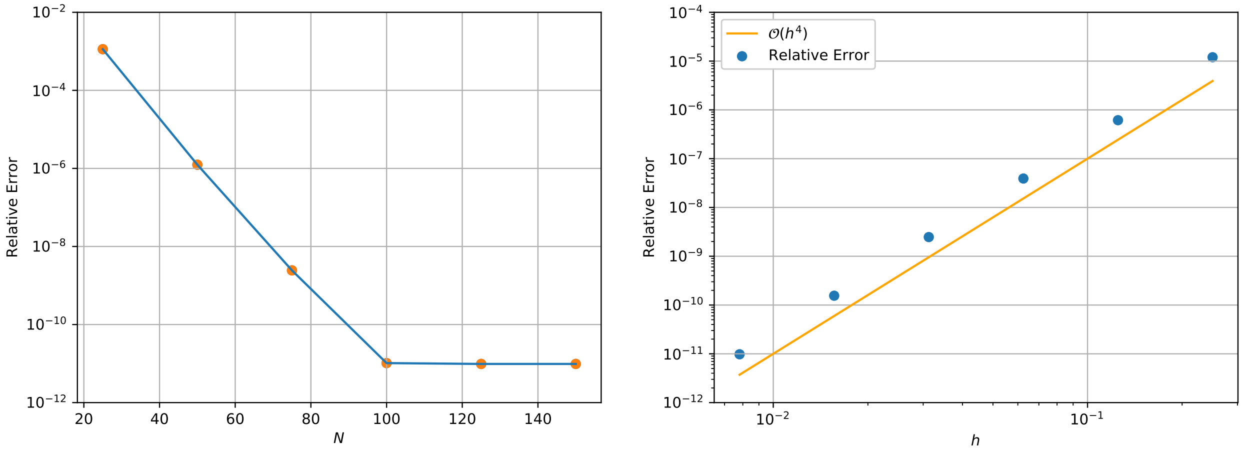


Figure 5.5: Spatial (*left*) and temporal (*right*) convergence at $T = 1$ for solving the Allen–Cahn equation using contour integral approach with quadrature nodes 2^6 on the contour to compute the φ functions of ETD4RK.

on the contour achieves the desirable accuracy. The numerical results confirm the accuracy with fourth-order temporal convergence and spectral convergence in space as shown Figure 5.5. Increasing the number of nodes for the quadrature rule in the contour integral improves the accuracy, as the error estimate in Eq. (4.21) predicts.

Experiment using the CF Method

We also implement the CF method to compute all the φ functions in the Krogstad [2005] form of EDT4RK. Solving the Allen–Cahn equation using rational approximation of type $(14, 14)$ leads to the fourth order convergence in time and spectral convergence in space as shown in Figure 5.6. It is worth mentioning that the number of poles used is far less than the number of nodes for the contour integral approximation.

5.2.2 Comparisons

We compare the complexity in the time-stepping or executions for the three methods used to compute the φ functions. Figure 5.7 shows the execution time in seconds

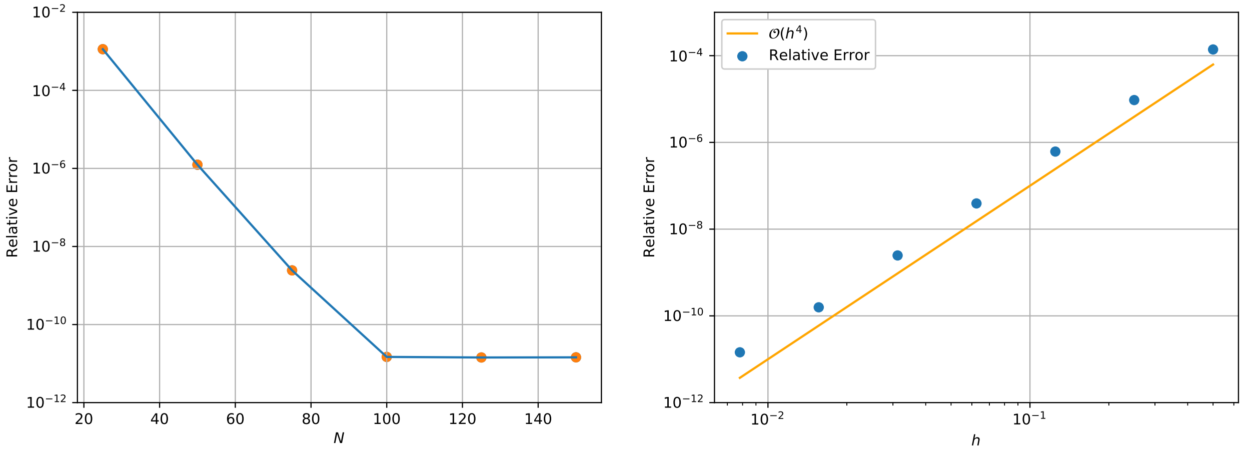


Figure 5.6: Spatial (*left*) and temporal (*right*) convergence solving solving the Allen–Cahn equation at $T = 1$ using type $(14, 14)$ CF approximants to compute the φ functions of ETD4RK.

against number of expansion coefficients of the solutions at final time $T = 1$. It does appear that it takes less than a second for the time-stepping using each of the three methods. The eigendecomposition approach is the fastest amongst the three methods. The rational approximation method seems to be more efficient than the contour integral approach due to the fact the number of poles required to achieve the desired accuracy is far smaller than the nodes in the contour integral approach.

The contour plots in Figure (5.8) show the number of nodes and poles needed for the contour integral and CF method respectively to achieve a desired accuracy. We do that in terms of the relative error in space against the number of spectral coefficients to represent the solution of the Allen–Cahn equation (1.7) simulated up to the final time $T = 1$. The error tends to decay relative to increasing the number of solution coefficients and the number of nodes for approximating the φ functions. We lose about 4 digits of accuracy for the CF method using a type $(16, 16)$ rational function and about 120 coefficients whiles about 6 digits loss of accuracy for the contour integral with 2^6 nodes and 100 spectral coefficients.

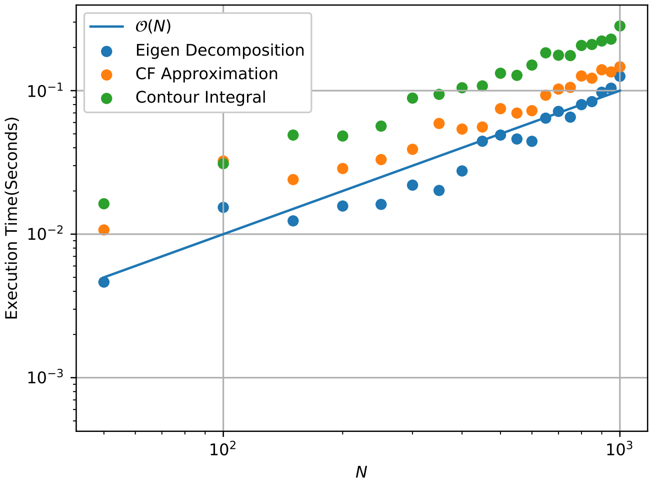


Figure 5.7: The execution time against the number of coefficients for the three methods used to compute the φ functions of ETD4RK for solving the Allen–Cahn equation (1.7) up to the final time $T = 1$. We use the type $(14, 14)$ for the rational approximation and 2^6 nodes for the contour integral.

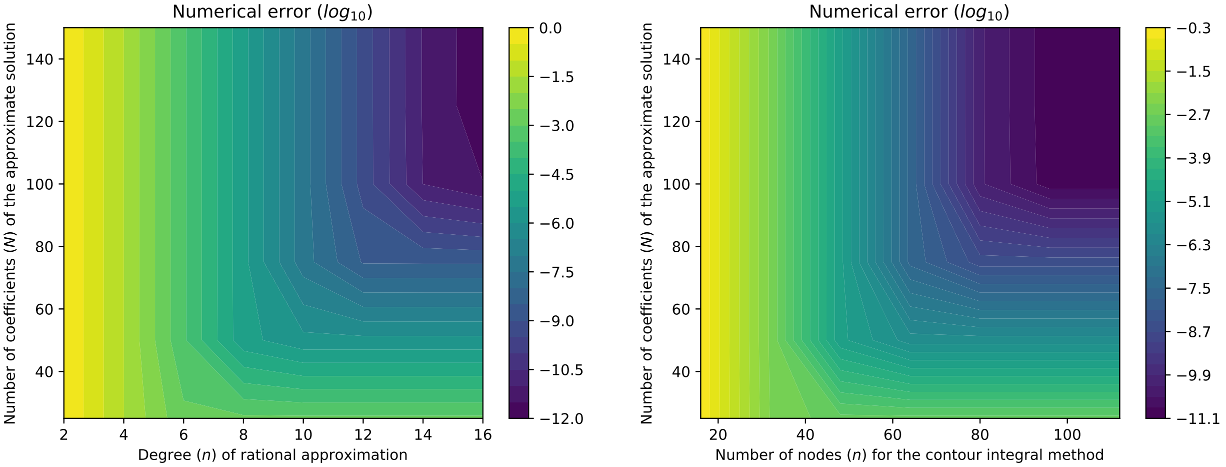


Figure 5.8: The contour plot of the error (log_{10}) at final time $T = 1$ for solving the Allen–Cahn equation using the CF (*left*) and contour integral (*right*) against the number of coefficients N of the solution.

The Kuramoto–Sivashinsky equation

We consider the Kuramoto–Sivashinsky (KS) equation given by

$$u_t = -u_{xxxx} - u_{xx} - uu_x, \quad x \in [-1, 1]. \tag{5.1}$$

The PDE (5.1) was derived by Kuramoto [1978] and Sivashinsky [1977] to model reaction-diffusion processes, specifically to study diffusive instability in a laminar flame front. It has second and fourth-order reaction terms and it is well known for producing chaotic behaviours (Lakestani and Dehghan [2012]). The negated second-order term $-u_{xx}$ destabilizes the system by producing energy which the nonlinear term uu_x transfers from low wavenumbers to high wavenumbers, while the fourth-order term u_{xxxx} term has a stabilizing effect.

We solve the PDE with homogeneous boundary conditions $u(\pm 1) = u'(\pm 1) = 0$ using the initial condition $u(x, t = 0) = \cos(x/16)(1 + \sin(1/16))$ (Montanelli and Bootland [2016]). Using the type $(14, 14)$ CF approximation to compute the φ functions in Eq. (2.18), the convergence results is shown in Figure 5.9.

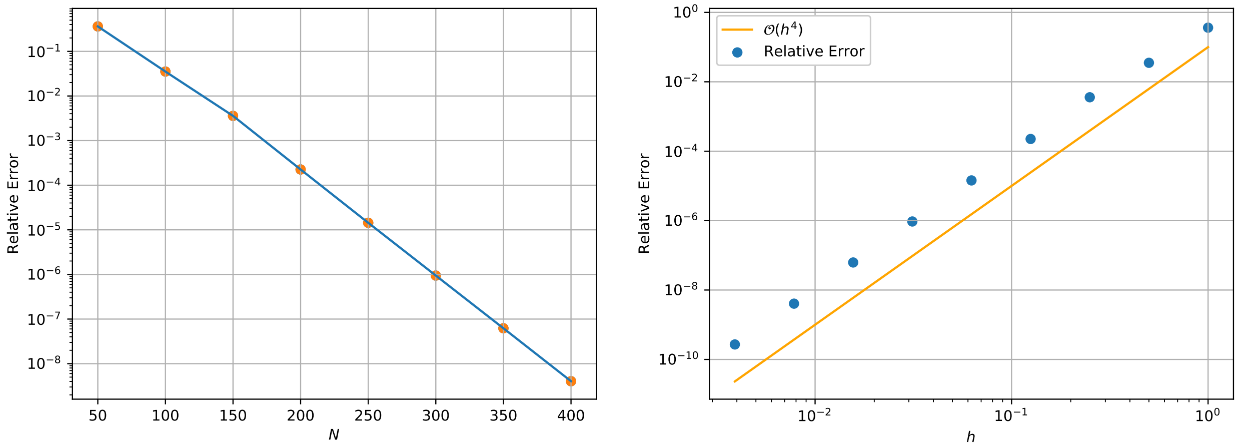


Figure 5.9: Spatial (*left*) and temporal (*right*) convergence of solving the KS equation at $T = 1$ using the type $(14, 14)$ CF approximation to compute the φ functions.

5.3 Nonhomogeneous Boundary Conditions

Consider the Allen–Cahn equation (1.7) with nonhomogenous Neumann boundary conditions $u'(-1) = -1$ and $u'(1) = 1$. We set $u(x,t) = v(x,t) + w(x)$ where $v(x,t)$ satisfies the homogeneous boundary conditions $v'(\pm 1) = 0$ and $w(x)$ the nonhomogeneous conditions $w'(-1) = -1$ and $w'(1) = 1$. From the basis recombination section 3.3, the choice of $w(x,t)$ based on the given boundary condition is given by $w(x,t) = 1/2x^2 - 1/6$ and thus Eq. (1.1) becomes

$$v_t = \mathcal{L}v + \underbrace{\mathcal{L}w + \mathcal{N}(v+w,t)}_{\tilde{\mathcal{N}}(v,t)}, \tag{5.2}$$

where $\mathcal{L}w$ is added to the modified nonlinearity $\tilde{\mathcal{N}}$. Thus we solve the homogeneous problem in v using the initial condition $v(x,0) = u(x,0) - w(x)$ and then use the relationship $u = v + w$ to get the solution in u. The convergence results are shown in Figure 5.10.

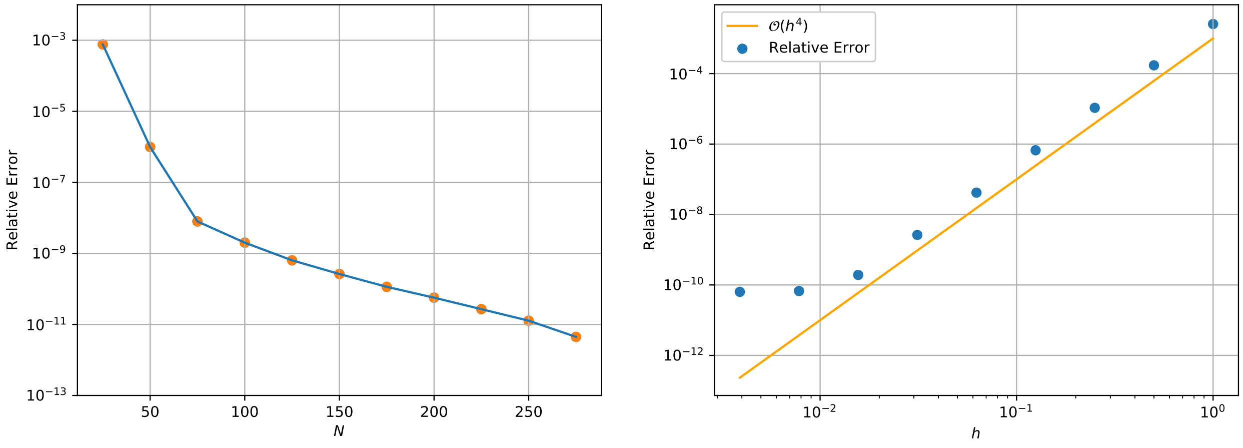


Figure 5.10: Spatial (*left*) and temporal (*right*) convergence for Allen–Cahn equation with nonhomogeneous Neumann boundary condition $u'(\pm 1) = \pm 1$ at $T = 1$.

5.4 Time-Dependent Boundary Conditions

Considering the heat equation with time dependent or non-constant forcing term Neumann boundary conditions as

$$u_t = \alpha u_{xx}, \quad \mathcal{B}u = u'(\pm 1) = c_\pm(t), \quad x \in [-1, 1], \quad t \geq 0, \tag{5.3}$$

where $\alpha = 10^{-3}$ is the diffusivity constant. For nonhomogeneous Neumann boundary conditions, we set the solution $u(x,t) = v(x,t) + w(x,t)$ such that $\mathcal{B}v = 0$ and $\mathcal{B}w = c_\pm(t)$. From the basis recombination in Section (3.3), the choice of $w(x,t)$ for this boundary condition is given by

$$\begin{aligned} w(x,t) &= c_-(t)\Big(\frac{1}{2}P_1 - \frac{1}{6}P_2\Big) + c_+(t)\Big(\frac{1}{2}P_1 + \frac{1}{6}P_2\Big) && (5.4) \\ &= c_-(t)\Big(\frac{1}{2}x - \frac{1}{12}(3x^2-1)\Big) + c_+(t)\Big(\frac{1}{2}x + \frac{1}{12}(3x^2-1)\Big) && (5.5) \end{aligned}$$

In this experiment, we choose $c_-(t) = \cos(2\pi t)$ and $c_+(t) = \sin(2\pi t)$. We solve the problem in v with the homogeneous boundary conditions and substituting $u = v + w$ in Eq. (5.3) leads to

$$v_t = \alpha v_{xx} + \alpha w_{xx} - w_t, \qquad v'(\pm 1) = 0, \tag{5.6}$$

where term $\alpha w_{xx} - w_t$ becomes the nonlinearity. The convergence results as shown in Figure 5.11 indicate a slow decay of the error in space, specifically with the error plateaus in space after $N = 300$ with error of 10^{-10}.

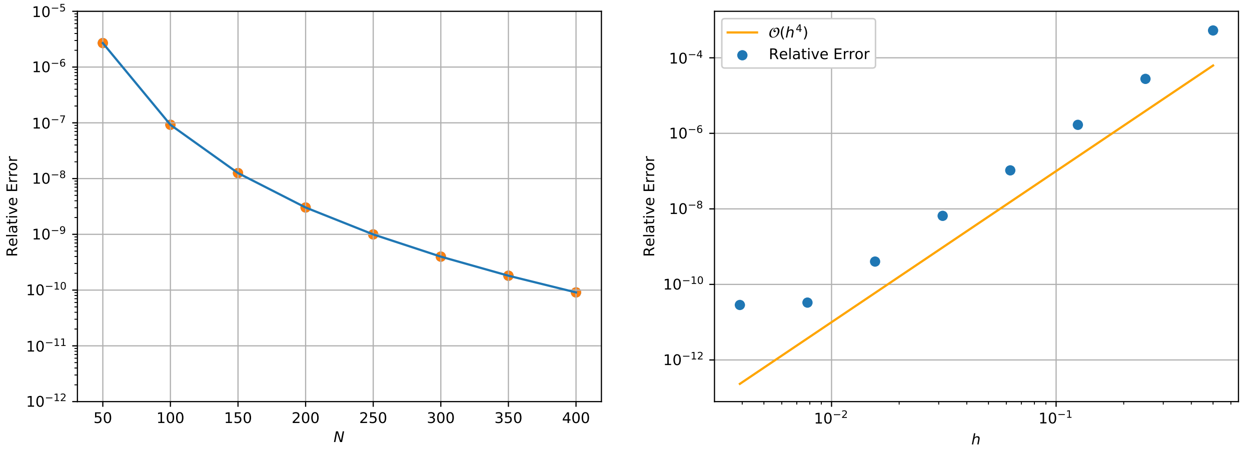


Figure 5.11: Temporal (*left*) and spatial (*right*) convergence of solving the heat equation (5.3) with non-constant forcing terms at the boundaries at final time $T = 1$.

5.5 Piecewise-Defined Problems

We consider the nonlinear Schrödinger-like equation with a piecewise-defined potential given by

$$i\epsilon u_t = -\frac{i}{2}\epsilon^2 u_{xx} + i|x|u + \epsilon|u|^2 u, \qquad x \in [-1, 1] \tag{5.7}$$

with the boundary conditions $u(\pm 1) = 0$. Eq. (5.7) can also be written as

$$u_t = \frac{\epsilon}{2}(u_{xx} - \frac{2}{\epsilon^2}|x|u) - i|u|^2 u, \tag{5.8}$$

where we treat $\frac{\epsilon}{2}(u_{xx} - \frac{2}{\epsilon^2}|x|u)$ as the linear part and the $i|u|^2u$ the nonlinear. It has a number of applications in physics including modelling the behaviour of quantum mechanical systems and nonlinear propagation of light in fibre optics. We modified the original equation so that the linear operator has real spectrum.

The convergence results, as shown in Figure (5.12), does not look as accurate as the other previous results. In particular, the fourth-order convergence in time is achieved with a specific range of the time step, after which the error plateaus and also very slow convergence in space.

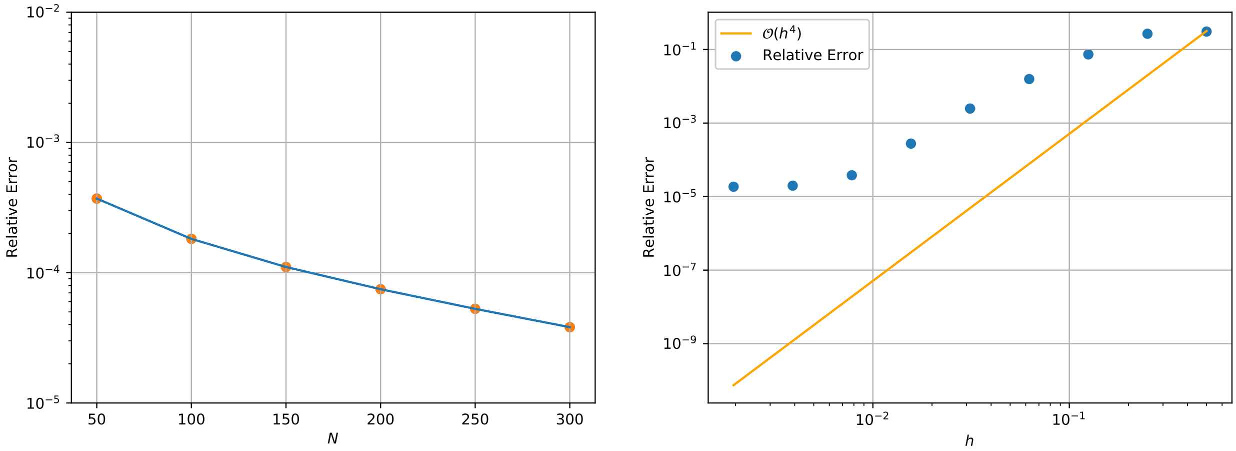


Figure 5.12: Temporal (*left*) and spatial (*right*) convergence of solving the nonlinear Schrödinger equation (5.7) at final time $T = 1$.

6

Conclusion

We have studied the numerical solution of stiff time-dependent semilinear PDEs with a self-adjoint linear differential operator for the linear part. The algorithm is based on the fact that the discretization of the self-adjoint operator using the structure preserving spectral method of Aurentz and Slevinsky [2020] leads to symmetric-definite and banded discretizations. The approach symmetrizes the ultraspherical spectral method using basis recombination. The symmetry guarantees real spectrum for any principal finite section and thus algorithms, based on the prior knowledge of the true spectrum, are developed to solve the resulting semi-discretizations.

We used exponential integrators which alleviate stability constraint of the step size due to the stiffness in the linear operator. The subsequent investigation of eigendecomposition, contour integral and Carathéodory–Fejér methods to approximate matrix-vector product involving the matrix φ functions led to a significant improvement in the complexity of the time-stepping scheme.

The algorithm is implemented to solve the Allen–Cahn, Kuramoto–Sivashinsky, heat and Schrödinger-like equations. Different boundary conditions including time-dependent and non-homogeneous boundary conditions were considered as well as piecewise-defined problems. The convergence results confirm the accuracy of the time-stepping and the spectral method.

This work can be extended in a number of ways. Possible future directions include: extending the methods for higher spatial dimensions; considering more examples of PDEs of this nature can lead to making the algorithm more general and possibly developing a numerical package for solving PDEs with self-adjoint linear differential operators; and, considering PDEs with complex spectra, such as skew-hermitian operators.

Bibliography

M. Abramowitz and I. A. Stegun. Handbook of mathematical functions. *US Department of Commerce*, 10, 1972.

B. K. Alpert and V. Rokhlin. A fast algorithm for the evaluation of legendre expansions. *SIAM Journal on Scientific and Statistical Computing*, 12(1):158–179, 1991.

J. L. Aurentz and R. M. Slevinsky. On symmetrizing the ultraspherical spectral method for self-adjoint problems. *Journal of Computational Physics*, 410:109383–1–24, 2020.

G. Baszenski and M. Tasche. Fast polynomial multiplication and convolutions related to the discrete cosine transform. *Linear Algebra and its Applications*, 252(1-3):1–25, 1997.

S. Bochner. Über sturm-liouvillesche polynomsysteme. *Mathematische Zeitschrift*, 29(1):730–736, 1929.

W. E. Boyce and R. C. DiPrima. Differential equations elementary and boundary value problems. *Willey & Sons*, 1977.

J. W. Cahn and S. M. Allen. A microscopic theory for domain wall motion and its experimental verification in fe-al alloy domain growth kinetics. *Le Journal de Physique Colloques*, 38(C7):C7–51, 1977.

L. Carlitz. The product of two ultraspherical polynomials. *Glasgow Mathematical Journal*, 5(2):76–79, 1961.

S. M. Cox and P. C. Matthews. Exponential time differencing for stiff systems. *Journal of Computational Physics*, 176(2):430–455, 2002.

G. G. Dahlquist. A special stability problem for linear multistep methods. *BIT Numerical Mathematics*, 3(1):27–43, 1963.

DLMF. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.27 of 2020-06-15. URL `http://dlmf.nist.gov/`. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

D. H. Griffel. *Applied functional analysis*. Courier Corporation, 2002.

W. Hahn. Über die jacobischen polynome und zwei verwandte polynomklassen. *Mathematische Zeitschrift*, 39(1):634–638, 1935.

N. Hale and A. Townsend. A fast, simple, and stable chebyshev–legendre transform using an asymptotic formula. *SIAM Journal on Scientific Computing*, 36(1):A148–A167, 2014.

N. Hale and J.A.C. Weideman. Contour integral solution of elliptic PDEs in cylindrical domains. *SIAM Journal on Scientific Computing*, 37(6):A2630–A2655, 2015.

N. Hale, N. J Higham, and L. N. Trefethen. Computing aˆα,\log(a), and related matrix functions by contour integrals. *SIAM Journal on Numerical Analysis*, 46 (5):2505–2523, 2008.

B. C. Hall. *Quantum theory for mathematicians*, volume 267. Springer, 2013.

J. S. Hesthaven, S. Gottlieb, and D. Gottlieb. *Spectral methods for time-dependent problems*, volume 21. Cambridge University Press, 2007.

N. J. Higham. *Accuracy and stability of numerical algorithms*, volume 80. SIAM, 2002.

N. J. Higham. *Functions of matrices: theory and computation.* SIAM, 2008.

M. Hochbruck and A. Ostermann. Explicit exponential Runge–Kutta methods for semilinear parabolic problems. *SIAM Journal on Numerical Analysis*, 43(3):1069–1090, 2005.

M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numerica*, 19: 209–286, 2010.

A. Iserles. *A first course in the numerical analysis of differential equations.* Number 44. Cambridge university press, 2009.

A. Kassam and L. N. Trefethen. Fourth-order time-stepping for stiff PDEs. *SIAM Journal on Scientific Computing*, 26(4):1214–1233, 2005.

H. L. Krall. On derivatives of orthogonal polynomials. ii. *Bulletin of the American Mathematical Society*, 47(4):261–264, 1941.

S. Krogstad. Generalized integrating factor methods for stiff PDEs. *Journal of Computational Physics*, 203(1):72–88, 2005.

A. Kufner. *Weighted Sobolev spaces*, volume 31. John Wiley & Sons Incorporated, 1985.

Y. Kuramoto. Diffusion-induced chaos in reaction systems. *Progress of Theoretical Physics Supplement*, 64:346–367, 1978.

M. Lakestani and M. Dehghan. Numerical solutions of the generalized Kuramoto–Sivashinsky equation using B-spline functions. *Applied Mathematical Modelling*, 36(2):605–617, 2012.

J. D. Lawson. Generalized Runge–Kutta processes for stable systems with large Lipschitz constants. *SIAM Journal on Numerical Analysis*, 4(3):372–380, 1967.

S. H. Lui. *Numerical analysis of partial differential equations*, volume 102. John Wiley & Sons, 2012.

A. P. Magnus. Asymptotics and super asymptotics for best rational approximation error norms to the exponential function (the '1/9'problem) by the Carathéodory–Fejér method. In *Nonlinear Numerical Methods and Rational Approximation II*, pages 173–185. Springer, 1994.

E. Martensen. Zur numerischen auswertung uneigentlicher integrale. *Z. Angew. Math. Mech*, 48:T83–T85, 1968.

B. V. Minchev and W. Wright. A review of exponential integrators for first order semi-linear problems. Technical report, 2005.

H. Montanelli and N. Bootland. Solving periodic semilinear stiff PDEs in 1D, 2D and 3D with exponential integrators. *arXiv preprint arXiv:1604.08900*, 2016.

H. Montanelli and N. L. Trefethen. SPIN, SPIN2, SPIN and SPINSPHERE for stiff PDEs, 2017. URL `http://www.chebfun.org/docs/guide/guide19.html`.

S. Olver and A. Townsend. A fast and well-conditioned spectral method. *SIAM Review*, 55(3):462–489, 2013.

A. Ostermann, M. Thalhammer, and W.M. Wright. A class of explicit exponential general linear methods. *BIT Numerical Mathematics*, 46(2):409–431, 2006.

M. Riesz. Sur le problème des moments, troisieme note. *Ark. Mat. Fys*, 16:1–52, 1923.

G. S. Sivashinsky. Nonlinear analysis of hydrodynamic instability in laminar flames—I. Derivation of basic equations. *In Dynamics of Curved Fronts*, 64:459–488, 1977.

M. Stone and P. Goldbart. *Mathematics for physics: a guided tour for graduate students*. Cambridge University Press, 2009.

A. Talbot. The accurate numerical inversion of Laplace transforms. *IMA Journal of Applied Mathematics*, 23(1):97–120, 1979.

L. N. Trefethen. *Spectral methods in MATLAB*, volume 10. SIAM, 2000.

L. N. Trefethen. *Approximation Theory and Approximation Practice*, volume 128. SIAM, 2013.

L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: the Behavior of Non-normal Matrices and Operators*. Princeton University Press, 2005.

L. N. Trefethen and M. H. Gutknecht. The Carathéodory–Fejér method for real rational approximation. *SIAM Journal on Numerical Analysis*, 20(2):420–436, 1983.

L. N. Trefethen and T. Schmelzer. Evaluating matrix functions for exponential integrators via Carathéodory–Fejér approximation and contour integrals. *Electronic Transactions on Numerical Analysis*, 29:1–18, 2007.

L. N. Trefethen and J. A. C. Weideman. The exponentially convergent trapezoidal rule. *SIAM Review*, 56(3):385–458, 2014.

L. N. Trefethen, J. A. C. Weideman, and T. Schmelzer. Talbot quadratures and rational approximations. *BIT Numerical Mathematics*, 46(3):653–670, 2006.

C. F. Van Loan and G. H. Golub. *Matrix computations*. Johns Hopkins University Press Baltimore, 1983.

J. Weideman and L. Trefethen. Parabolic and hyperbolic contours for computing the Bromwich integral. *Mathematics of Computation*, 76(259):1341–1356, 2007.

H. Wilbraham. On a certain periodic function. *The Cambridge and Dublin Mathematical Journal*, 3:198–201, 1848.